UNCLASSIFIED
Security Classification This Page

REPORT DOCUMENTATION PAGE

1. Report Security Classification: UNCLASSIFIED

2. Security Classification Authority: N/A

3. Declassification/Downgrading Schedule: N/A

4. Distribution/Availability of Report: UNLIMITED

5. Name of Performing Organization: ADVANCED RESEARCH DEPARTMENT

6. Office Symbol: 1C

7. Address: NAVAL WAR COLLEGE, 686 CUSHING RD.,
NEWPORT, RI 02841-5010

8. Title (Include Security Classification):

MILITARY APPLICATIONS FOR INFORMATION EXTRACTION TECHNOLOGY

9. Personal Authors: Paul S. Walczak, Major, USA

10. Type of Report: Final

11. Date of Report: 16 June 95

12. Page Count: ~~85~~ 84

13. Supplementary Notation: A paper submitted to the Faculty of the Naval War College in partial satisfaction of the requirements of the Department of Advanced Research. The contents of this paper reflect my own personal views and are not necessarily endorsed by the Naval War College or the Department of the Navy.

14. Ten key words that relate to your paper:
Information Extraction, Information Operations, Information Warfare, Advanced Information Processes, Hyperspace, Anti-Information, Systems Automation, Natural Language Processing

15. Abstract: The amount of information that must be processed for military purposes presents a daunting challenge to administrators, analysts, and leaders. The advent of information extraction technology that uses artificial intelligence techniques to apply meaning to data is a current research effort being funded by DoD. While information extraction has much potential to be employed in military tasks, little has been done to identify specific military applications for this technology. This paper describes information extraction and suggests several possible uses across a broad range of military interests.

16. Distribution / Availability of Abstract: A	Unclassified	Unclassified	DTIC Users

18. Abstract Security Classification: UNCLASSIFIED

19. Name of Responsible Individual: Chairman, Department of Advanced Research

20. Telephone: (401) 841-3304

21. Office Symbol: 1C

Security Classification of This Page UNCLASSIFIED

NAVAL WAR COLLEGE
Newport, R.I.

MILITARY APPLICATIONS FOR INFORMATION EXTRACTION TECHNOLOGY

by

Paul S. Walczak

Major, USA

A paper submitted to the Director, Advanced Research Department, as an Advanced Research Project in partial satisfaction of the academic requirements of the Naval War College for the degree of Master of Arts in National Security and Strategic Studies.

The contents of this paper reflect my own personal views and are not necessarily endorsed by the Naval War College or the Department of the Navy.

Signature: Paul S Walczak

16 June 1995

Paper directed by
E. Nielsen, CAPT, USN
Chairman, Department of C4I

EXECUTIVE SUMMARY

MILITARY APPLICATIONS FOR INFORMATION EXTRACTION TECHNOLOGY

Information extraction tasks require the analyst to identify, transcribe, and organize relevant information from raw data for use in intelligent processes. Today's information requirements however, pose a vexing dilemma to analysts and decision-makers who rely on such extracted information. For example, there is more data available than can be effectively processed into useful information. To solve this inadequacy, researchers are developing automated information extraction technology. Information extraction systems, based on artificial intelligence theory, may soon be capable of performing the most tedious of data processing tasks, significantly enhancing the ability to manage relevant information.

Military exploitation of IE technology involves influencing three paradigms that bear on the problem: military intelligence; information science; and natural language processing. Together, these disciplines comprise the pertinent technical background which must be understood to attain a broader assimilation of IE technology for military purposes.

The military applications that are proposed here serve only to demonstrate the use of IE to solve a broad range of military information problems, from personnel management and health services, to military intelligence and command and control issues. Potential threats, which arise from the leveraging of IE capability, are suggested. Vulnerabilities, posed by the abundance readily available information in American society, are also examined.

The futuristic concept for a "hyperspace" suggests further direction for expanding the potential for IE technology. The hyperspace inter-links information requirements (represented by IE templates) across organizational boundaries as an enhancement to the current practices of simply inter-networking communications.

PREFACE

I discovered information extraction technology as a potential research area while trying to find topic information for another Naval War College assignment. Through a series of queries using the World Wide Web intelligent search tool on the Internet, I stumbled across a document authored by Professor Wendy Lehnert of the University of Massachusetts. Her brief article summarized the subject of information extraction, and indicated that DoD funding was the principle source for continuing development of this emerging field. I became curious as to the nature of this topic since my next military duty will be in a related career field, and because the military application for this technology was not readily apparent.

After conducting initial research for this topic, I discovered that the potential IE held for military applications, outside of the intelligence specialty, had received little attention. Additionally during this phase of the project, I was not able to find a military officer who could describe what information extraction was. These facts helped me to develop a thesis for undertaking this academic project:

> Examining the nature of information extraction technology can provide greater insight into both IE's potential to serve the military mission, and the state of the military information architecture.

My objective for this project was to provide a general, technical background in the subject, and to present my own ideas for developing information extraction technology into potential military applications. I have sequentially arranged this paper to: (1) relate IE to common military concepts; (2) describe the technology underlying information extraction; (3) identify applications that exploit IE for military purposes; and (4) examine possible threats to national interests that may be presented by IE.

The introductory chapter focuses on relating information extraction technology to issues relevant to today's military. Here, I refer principally to military documents in order to tie information extraction technology to military principles. The scale of existing information problems related to national security is described in terms of size and complexity. The second chapter discusses IE as a model under each of its inherent disciplines. The third chapter explains IE at a level that does not require technical familiarity.   The fourth chapter is a brief summary of IE research and development programs. The fifth chapter presents military application areas that I propose for exploitation by IE methods. The sixth chapter outlines possible threats and vulnerabilities that may arise from the proliferation of IE technology. The seventh chapter examines model-based extraction methods and proposes the concept for a "hyperspace".

Further Research Opportunities

I originally intended on presenting a broader report on the capabilities and potential for information extraction technology, to include an expanded analysis of imagery extraction. However, that scope quickly became untenable, and I was forced to focus almost entirely on text-based IE. I believe that the following areas hold rich opportunity for future research projects similar in nature to the Naval War College's Advanced Research Program:

* Image Interpretation/Imagery Understanding
* Information Retrieval
* Message Understanding (deeper textual processing and analysis than IE attempts)

* Integration of Advanced Information Processing and Analysis Technology (hyperspace)

* Speech Recognition Applications for the Military

* Document Management Applications for Intelligence

* Military Information Systems Performance Metrics

* Review of Military Doctrine for Information and Intelligence Support to Operations

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Information extraction (IE) is the basic skill utilized by the researcher and analyst. It is the process that recognizes and organizes information from within large sets of data. The objective for information extraction tasks is to make information available to information-dependent applications.[1] Contemporary research and development, conducted using defense resources, demonstrates automated IE capability by using computers to extract information.[2] The topic of this paper concerns the automation of information extraction and how IE may be exploited for military purposes.

Doctrinal Foundations

The most important application for information extraction is as a tool in support of information warfare. The current National Military Strategy emphasizes the criticality of information within the principle for winning the "Information War":

> The remarkable leverage attainable from modern reconnaissance, intelligence collection and analysis, and high speed processing and transmission warrants special emphasis. The services and combatant commands require such fused information systems. These systems enhance our ability to dominate warfare.[3]

[1] Information-dependent applications are those that need specific information to complete their processes. Synonymous to "knowledge-based applications" and "intelligent processes" within the scope of this paper.

[2] "Computers" are referred to in the research literature as "machines", a distinction borrowed from the field of artificial intelligence.

[3] Joint Chiefs of Staff, National Military Strategy of the United States of America, A Strategy of Flexible and Selective Engagement , February 1995, pg. 15.

To achieve U.S. dominance in information warfare, a doctrine for Information Operations (IO) is being developed which creates an architecture for providing information support to military operations under this guiding principle. Information operations are continuous operations that enable, enhance, and protect the commander's decision cycle and mission execution while negatively influencing an opponent's.[4] IO is an overarching concept for the inter-disciplinary exploitation of information that provides structure and organization to the way the American military will conduct information warfare.

IE can be used to leverage information resources in several different military disciplines as a specific requirement of information operations. IE technology, can affect each focus area (Intelligence, Independent Media, Friendly Capabilities, and Military Information Warfare) that comprises the information operations model. For example, the automated extraction of information from data can be employed to support psychological operations (PSYOP) objectives. These PSYOP objectives have related intelligence tasks, which may employ IE on products from the independent media. The use of IE to process text generated by the media requires the harnessing of the processing capability of friendly computing systems. A broader discussion of potential IE applications appears in chapter 5 of this paper.

Shortage of Resources, Increased Complexity, Excess Data

Shrinking resources, combined with the mounting tide of unused information, cause concerns about the quality, and even the rationality, of decision-making. The political impetus for reducing the military's budget promotes the search for new information processing techniques that may cut the tail-to-tooth ratio (the

[4] U.S. Army Dept. Information Operations (Coordinating Draft), FM 100-6 , n.p. July 1994.

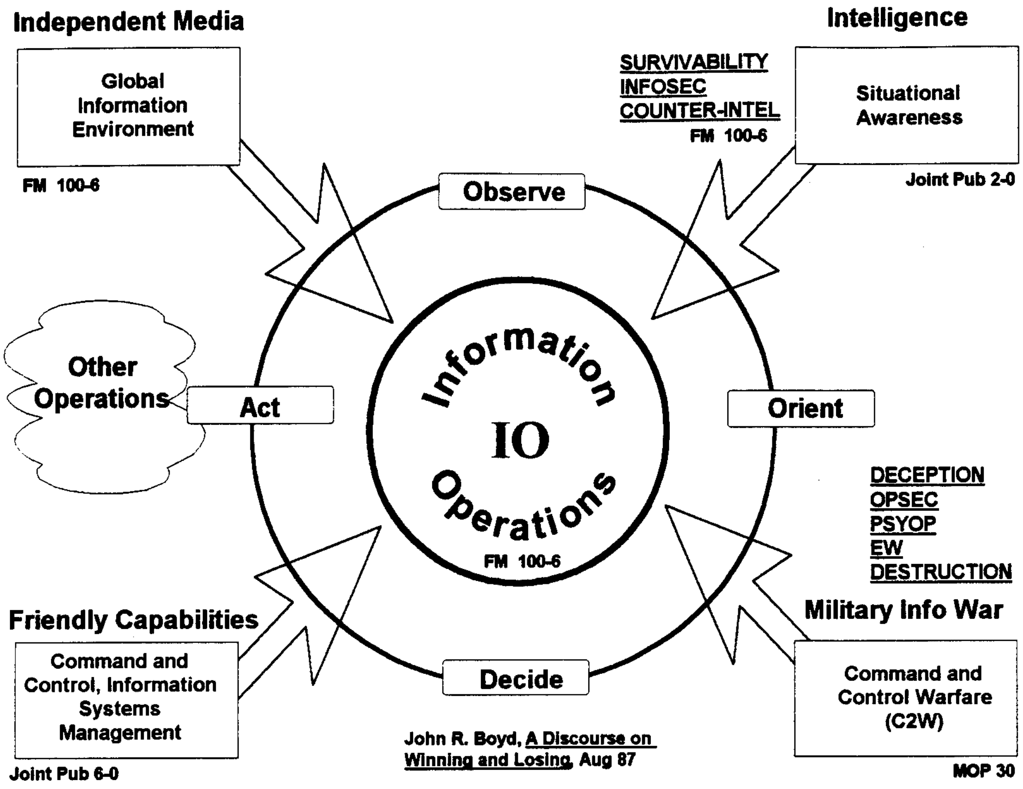


Figure 1 - Information Operations

amount of support it takes relative to the actual combat-capable force).[5] Automation of IE tasks may help cut overhead costs from the budget, and improve the efficiency with which information is processed.

The sea of data that exists today is a daunting challenge to decision making processes. For example, during the Gulf War, 8 gigabytes of information were processed each day by the Joint Intelligence Center (JIC). Had all of this information been in text form, this figure would represent the equivalent of 4 million

[5] Admiral Owens, Vice-Chief of the Joint Chiefs of Staff, alluded to these points, emphasizing the keen interest the U.S. military must have concerning the exploitation of information technology during an address to the Naval War College body on April 28, 1995. See also references to Patrick Wilson's work on "Unused Relevant Information in research and Development", under this paper's discussion of Anti-Information.

printed pages. The human resources required to adequately (in terms of speed and accuracy) sort through this amount of information manually is infeasible. Cause for additional concern is the anticipated growth rate of operational information flow that is expected to reach 100 gigabytes per day by the turn of the century and the impact that "digitizing the battlefield" will have on the processing requirement.[6] Advanced information processing capability will be crucially needed at the military's operational level.

Equally significant to the increased need for improved information processing is the tremendous political and cultural changes which have been witnessed recently throughout the world. The broad geo-political restructuring in this era has been accompanied by both, an ensuing volatility in the representation of "truth", as well as increases in factors of information supply and demand.

The addition of new sources of information adds to the supply of potentially relevant information. Previously, nations with authoritarian governments communicated to the international community using a single, unequivocal message, often through a single medium. These same societies are now able to fully exploit today's modern communications systems. For example, Turkey, a nation at the geographic apex of potential global tensions, has witnessed remarkable growth in information opportunities. Since the early part of this decade, more than 500 television stations, and 2,000 radio stations been introduced across this nation of 60 million people. Prior to this boom, only five television channels were accessible in the nation. Many of Turkey's most recent broadcast venues have specific audiences as targets, such as Islamic fundamentalists, and the Kurdish minority, while others represent specific phenomena creates the opportunity to observe situations from many diverse

[6] Department of the Army, Training and Doctrine Command, "Force XXI Operations". TP 525-5. Digitization of the battlefield, or providing better opportunity for the commander to know what the situation is within his battlespace area of interest is a major theme of the Army's Force XXI project.

political interests, such as those of the Welfare party.[7] This communication perspectives.

Quicker reporting on the feedback to changing conditions, through foreign or domestic media sources, means more data to analyze. Our ability to understand the nature of events may be limited by conceptual complexity, resulting from the cultural and political implications of rapid information exchange, as well as by a shortfall in information processing capability. The current situation in Bosnia is a good case in point. The impact of immediate media coverage of events, such as the taking of United Nations personnel as hostages or the downing of an American F-16 jet, influences the political and military responses of the parties involved. Hence, we will likely experience a decline in both our ability to interpret volatile situations as well as in our willingness to react decisively toward them.

To conclude, information extraction technology and its relevance to military application can be framed using this familiar context taken from the military operational thought process:

Ends: Support the attainment of military objectives through the ability to process more data into relevant information, faster, with less resources.

Ways: Automated processing and analysis capabilities for decision making and command and control

Means: Information Extraction Technologies

[7] John Pomfret, "Islam Takes Message To Turkish Television", Washington Post, 19 April 1995, pg. A 25.

CHAPTER 2

CONCEPTS RELATED TO THE INTEGRATION OF INFORMATION EXTRACTION

If we consider a single piece of information as a "fact," then information extraction is concerned with retrieving facts from data. A fact may be considered as an aggregate of individual data items. A fact could also be derived from imagery data, such as the existence of an object at a certain place and time.

IE may be employed to find data that has distinct orientation for different purposes within an information architecture. Davidow and Malone provide an interpretation of four contexts of information in their discussion about the power of useful information in "The Virtual Corporation":[8]

> At the lowest level, content information is concerned with quantity, location, and types of entities. Content is historical in nature, and is the most likely kind of information to be found in an organization's database.
>
> Form information describes the shape of an object.
>
> Behavior information is associated with prototyping, modelling and simulation of the effects that the environment has on particular objects.
>
> Action information is used to effect an outcome. It is information that instantly converts to a sophisticated action. Action information is the information which feeds artificial intelligence applications.

IE systems may find data that fits each of these categories. Content data can be extracted from fiscal reports, country profiles, historical archives and individual

[8] William H. Davidow & Michael Malone, The Virtual Corporation-Structuring and Revitalizing the Corporation for the 21st Century, New York: Harper Collins, 1992. pg 67.

bank statements. Image sources, and even narrative text, provide information. The information of most interest to national security, however, comes from the behavioral category. For example, in the case of behavioral modelling, the subjects are human entities, such as terrorists, political parties, or extremist organizations. Action information also is directly related to IE capability; it is information that is required for "sensor-to-shooter" concepts, which envision the activation of weapons platforms based on near real-time processing of raw data collected from disparate sensors.[9]

Data

Datum placed into a data structure, is given an explicit information value based on the criteria it has met. However, it has not been converted into useful, or relevant, information until a user, or an application, accesses the data and aggregates or computes from the dataset. Once information has been compiled, it can be interpreted by the consumer (the human or machine user of information) to produce "knowledge", or in terms related to the interpretation of information regarding military adversaries, "intelligence".

After data has been organized into structures within a database, queries can be constructed to retrieve information to satisfy a user's needs. This capability assumes that the user understands the basic task in manipulating a database, and is experienced in the database system's query language or interface.

The results which are obtained through traditional queries are generally representative of set classification tasks established by the user. For example,

[9] "Real-time" is considered to be almost instantaneous. The only delay between an observation and a reaction would be the time delay in processing the observation and notifying the weapons components to fire. Near-real time adds in the communication delay time that is required between the sensor, the processor, and the weapon component. The major characteristic is that humans may not be part of the decision cycle.

elements of the database having discrete characteristics identified by the user, are grouped together in a resultant set (these matching elements are known as "hits" in database management vernacular) for further inspection, editing or computation, such as summation of the data in record fields, or other statistical analysis. Database management systems, (DBMS), usually incorporate a robust suite of analytical tools that can be applied in conjunction with the query itself. These capabilities are often customized locally in applications adapted to an organization's needs, or acquired in the form of an enhancement package to a specific DBMS.

This traditional approach to retrieving information confines data to follow deliberate sequences of logic, and results in sets of data whose attributes match those desired in the query. Generating sophisticated querying techniques for existing databases is another approach to information extraction. These techniques have the potential to extend the capability of the database application's query language into meaningful "hunts" for information, conceived in terms related to natural perception.

Concepts

A "concept" is defined as a generalized idea of a class of objects; a psycho linguistic category that can be either physical (cold or tank) or abstract (morale or strategy).[10] Humans abstract meaning from sensory data based on relationships they associate to that data. A concept could help define an object, such as the visual form that distinguishes a personnel carrier on the battlefield, or it could result in a judgmental conclusion, based on the interconnection of diverse data elements.

[10] Definition taken from Webster's New World Dictionary, 2nd College Edition. (NY: Simon & Schuster) 1980, and Los Alamos National Laboratory, "Data Mining for Concept Extraction." Available through the Internet's WorldWide Web information retrieval tool at the Lab's "home-page" address of http://www.lanl.gov.

Anti-Information

The term "anti-information" is derived from Patrick Wilson's study on the nonuse of information, and the work of Richard K. Betts on the subject of war and intelligence failure.[11] The work of these authors characterizes anti-information as either relevant information known to exist but not used, or ambiguous information that decrements a knowledge process by creating greater uncertainty. The existence of anti-information provides compelling reasons for pursuing information extraction technology because IE may eliminate anti-information as a concern.

Uncertainty and Overload. In "Analysis, War and Decision: Why Intelligence Failures Are Inevitable,"[12] Betts discusses the "ambiguity of evidence." According to Betts the role of intelligence is "to extract certainty from uncertainty and to facilitate coherent decision in an incoherent environment." He further describes the dilemma of trying to reduce uncertainty by extracting information from "evidence riddled with ambiguities," and the risks professional analysts are subjected to when they "oversimplify reality and desensitize consumers of intelligence to the dangers that lurk within the ambiguities." In these cases it is often assumed that a state of uncertainty suggests the lack of information. But, as Betts continues, "ambiguity can also be aggravated by an excess of data." The term "information overload" is understood to mean having more data than one can manage. From Bett's perspective, documents presented to decision makers lacked appropriate levels of objectivity and rationality, as a consequence of information overload.

[11] Patrick Wilson, Unused Relevant Information in Research and Development, Journal of the American Society of Information Science, volume 46, no. 1. 45-51.

[12] Reprinted from World Politics, vol XXXI, No. 1, October 1978, (included in materials issued for the NWC elective program). pg. 41.

Rational Decisions, Overload and Deliberate Non-Use. Wilson hypothesizes that the efficiency of decision-support systems, whether stock markets or military organizations, is based upon rationality. Rationality requires the use of all relevant information in arriving at theoretical and practical solutions.

Wilson also discusses "overload" and deliberate categories of non-use that challenge rationality.[13] He defines the '"serious kind of overload" as the "possession, or knowledge of the existence of information one thinks to be probably relevant but does not use due to lack of time." A backlog is created once all available time is committed to more urgent activities than the processing of information. The backlog continues to grow until parts of the backlog are never used. Wilson states that it is "a plausible hypothesis that overload is characteristically associated with accumulation of backlogs, and that the more severe the overload, the more likely it is that portions of the backlog will never be used."

Deliberate nonuse of relevant information occurs in two ways: information might simply be ignored, or it might be blocked by an assumption. According to Wilson, available information on a topic may be ignored in analysis for the following reasons:

Deferral - Consider information later (or not at all).
Specialization - Focus is on other aspects of the subject.
Territoriality - Parochialism; topic from another specialty
Safely ignorable - A topic is relatively unimportant.
Unmanageable - Topics exceed competence or resources.
Oversupply - Too much work for the analyst (backlog)[14]

An assumption may be defined as a temporary placeholder to be replaced by

[13] Other kinds of nonuse are outside of the concept of "anti-information": Assuming that analysts want to obtain all available relevant information, failure to find sources is regarded as non problematic, merely a mistake; Irrelevant or obsolescent information, is discarded.

[14] Wilson, pg. 47.

information made available in the future, or it may be adopted to make progress possible on an otherwise intractably difficult problem. Analysts may adopt an assumption as a substitute for gathering and using available information. Assumptions may pertain to only one dimension of a problem. For example, an analyst may work entirely with idealized assumptions, substituting the construction of idealized models for the direct investigation of reality.[15] If the assumption is not accurate, it becomes anti-information.

Betts and Wilson together provide a different perspective for the potential employment of automated information processing. Wilson concludes that information management calls for aids in screening, evaluating, and filtering, "not just to distinguish relevant from irrelevant, but to separate dispensable from indispensable relevant material." Applications for IE that seek to eliminate problematic issues in cases of information overload and the deliberate non-use of information in military settings should be explored.

Intelligence

Objectivity is a principle of intelligence production that has particular significance for IE.

> Analysts must be unbiased and avoid any tendency toward preconceived ideas.... If time or resources are inadequate to provide unambiguous intelligence, the JFC [Joint Forces Commanders] should be made aware of the ambiguity or uncertainty. Commanders need all available pertinent intelligence, including conflicting or contradicting information and opinion. Commanders may not be able to use or resolve conflicting intelligence, but it must be available for potential applicability as operations evolve and the actual situation becomes more apparent.[16]

IE may be employed to counter conditions where objectivity is questionable.

[15] Ibid.

[16] Joint Pub 2-0, Intelligence Support for Operations, pg. IV-21.

In its present state of evolution, IE has been established as an advanced information processing technique for "intelligence" purposes. Information extraction is a technical capability for information processing, that has evolved within the intelligence community. Information processing converts information into a form that can be used to produce knowledge, or intelligence, and includes tasks that translate data formats from several different media. Production is the integration, evaluation, analysis, and interpretation of information from collected sources that is accomplished to meet specific intelligence requirements.[17] The products of the intelligence cycle are intelligence applications. These applications, or products, are defined in the doctrinal intelligence publications as the "direct *extraction* and tailoring of information from an existing foundation of intelligence and near real-time reporting."[18] The fact that IE is so closely associated with military intelligence, and its focus therefore narrowed, may impede the full exploitation of IE across the varied military disciplines.

The parochial distinction between "information" and "intelligence" is established in Joint Pub 2:

> *Information* is data that have been collected but not further developed through analysis, interpretation, or correlation with other data and intelligence. The application of analysis transforms information into *intelligence*. Both information and intelligence are important, and both may exist together in some form. They are not, however, the same thing, and thus they have different connotations, applicability, and credibility.[19]

Military intelligence confines IE to a role in the production of intelligence (knowledge about an adversary) rather than for the extraction of broader information

[17] Ibid., pg II-7 & II-8.

[18] Ibid. pg. II-8. Examples of intelligence applications include intelligence preparation of the battlespace, mission planning, support to exercises, secure video-teleconferencing, and electronic mail (E-mail).

[19] Joint Pub 2-0, Intelligence Support for Operations, chapter II.

needs. Military information requirements are present in the wide variety of knowledge functions that must be accomplished across the range of military operations. Some of these requirements are for the sustaining base, while others occur at operational and tactical levels of war, in both support roles and combat arms disciplines.[20] The emerging doctrine for information operations, cited in the first section, holds more potential to provide a proper azimuth toward integrating multi-disciplined interests in IE development.

Natural Language Processing (NLP)

People communicate using language. A developed language is essential to the natural interaction of humans, and the growth of civilization. A "natural language" is what people normally use to communicate with others in their daily functioning.

Artificial intelligence (AI) is the area of scientific research that seeks to develop machines that replicate the human cognitive process. NLP is a discipline within the field of AI, and is the underlying technology on which automated information extraction technology is based. NLP uses incremental processes to derive meaning from the symbolic representation of language.

Natural Language. Written natural language is the notation of communication using a set of character symbols known as an alphabet. The symbols of an alphabet are grouped to form higher order representations of language called words. The words are arranged into sentences that convey independent meaning. Sentences which thread a common idea are collected into paragraphs. When these written language structures are assembled together to communicate a specific purpose, or

[20] The reference to combat arms disciplines is synonymous to the term "war fighter."

message, the result is referred to as a text.

Sentences contain both the ordinal structure and the meaning of the words. The sentence is the lowest-ordered language structure that communicates an idea, or concept. Sentences use words to define objects, or things (nouns) and words to describe actions (verbs) that involve those objects. Other words are used to link or orient actions and objects (i.e. - pronouns, prepositions, adjectives, adverbs and articles). The sequence with which sentences are structured is critical to be able to interpret the concept being communicated.

Machine Processing. In order for a machine to be able to comprehend written notation, the character symbols must come from a set that is understood by the machine. The ASCII character set is the common medium for computer-aided natural language understanding.[21] Speech understanding requires the capability to process human voice sources, and translate this auditory information into ASCII. Processing consideration for speech and written language are similar once they are both in this common text form.[22]

Text understanding requires the ability to accurately interpret: the meaning of individual words; the structural relationship of phrases and word sequences within a sentence; and the interrelationship of words and structures as abstractions of thought.[23] Progressing sequentially through these interpretative stages reduces the content of a text document from a symbolic representation of language into a

[21] American Standard Code Information Interchange - a "lowest common denominator" character representation for computer interpretation.

[22] Many printed alphabets and spoken tongues may appear in the potential information base. The first step in processing data originating in a foreign language is translation to a English, possibly employing machine translation tools.

All references to "text" imply converted speech data. Issues for the collection and conversion of speech to text format is outside the scope of this paper.

[23] Paul E. Lehner, Artificial Intelligence and National Defense- Opportunities and Challenge, Blue Ridge Summit, PA: Tab Professional and Reference Books, 21-34.

concept that is understood by a machine.

The Meaning of Words - Lexical Analysis. Machines use dictionaries, stored in digital form (or "digitized"), to look up the meaning of words. The problem with interpreting word definitions, even with the help of digitized dictionaries, is that words can have many meanings. A process called "lexical disambiguation" within the NLP development community, is used by computers to resolve meanings. Lexical disambiguation uses other understood parts within a sentence structure to determine the specific meaning of a word contained by that sentence. This process is similar to that used by human readers, more familiarly referred to as "taking meaning from context." Once word meanings have been interpreted, the grammatic structure, or model, of the next higher language object, the sentence, can be defined.

Structural Analysis. A grammar is a set of rules that defines how classes within a language may be structured to communicate meanings. A grammar is used to correctly interrelate words and phrases and is essential to sentence interpretation.

A sentence is the first major grammatical construct or "package." The term package, because that is essentially what a sentence is: an object confined in a boundary defined by a large capital letter at its beginning and a period at its end.

Structural analysis, or "syntactical analysis", is analogous to opening a box, disassembling the object contained in the box, and then inventorying the parts. "Reverse engineering" of a grammatic structure is analogous to "syntactic analysis"; e.g. the sentence is disassembled into its components to gain understanding of its construction.

Parsing is the commonly used term for describing the reverse engineering process that determines the structure of a sentence. Parsing uses the rules of grammar to establish valid relationships between components of a sentence. Parsing is also a modular breakdown sequence that implies levels of significance and corresponding relationships for language components. Several valid relationships may exist for a single term or between groups of terms. This fact makes automated parsing extremely complex since machines must be programmed to differentiate between several parse solutions to produce the most correct representation of a sentence. The product of parsing is a diagrammatic structure called a "parse tree." The root of the parse tree is the sentence; the first offspring are usually noun and verb phrases; parsing continues until the lowest level language component has been assigned a relationship along one of the parse tree's branches. Various authors use several approaches to characterize the parsing process.[24]

Interpreting Content. Semantic analysis renders conceptual understanding from sentences incremental: the values of the words contained by the sentence, when arranged in the structure derived through syntactic analysis, produce a clear, singular idea from the textual representation. The nuances of communication often cause the meaning, defined by relationship of the language terms, to be dependent on structural (syntactical) analysis. This interdependency causes some NLP systems to combine semantic and syntactic analysis in one step. Specialized techniques, such as creation of semantic and case grammars, and transformation grammars, have been employed to combine the abilities of semantic and syntactic analysis.[25]

[24] Ibid., pg 24.

[25] Gerard Salton, Dynamic Information and Library Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975, pg. 398., discusses transformation grammars.

Lehner, 26. Using semantic grammar, syntactic parsing methods can be employed to generate semantic parse trees. A disadvantage in utilizing semantic grammars is the large number of rules required to handle every possible relationship between terms relevant to the subject domain. Case grammars translate the results of a syntactic parse into a deeper, more semantically-oriented structure. There are very detailed and

The semantic processing of a text is one of the most important parts of NLP inside of IE systems. Each IE task is performed to extract specific kinds of information from documents. The semantic analysis stage of NLP for IE is somewhat dependent on the criteria used in filling data slots of information templates.

Discourse Processing. One of the major problems presented to a machine in NLP is that of co-reference resolution, or discourse processing. Discourse processing means recognizing references to already identified objects later in a text (the discourse of a topic). To achieve this ability you must associate amplifying or clarifying data related to an object that may be outside the initial reference in the text. An example of co-reference resolution through discourse processing is one where the same event is referenced within different sentences. The event may be described as a "war-crime" in one instance and later as a "massacre" in another. The machine must somehow be able to interpret that two references to the same event has occurred and that ensuing relevant data applies to the first (same) event, and should not cause the system to produce a new template.

Discourse processing represents a semantic problem in that a machine must be able to understand terms similar in meaning to a single concept, which implies "synonymity", and the use of a thesaurus. Beyond synonymity, the machine must be able to distinguish whether the terms in question actually represent distinct objects. These nuances are critical because there are many different possibilities for correct grammatical interpretations. Even though the grammatical interpretation may appear correct, the conceptual associations may be incorrect. The ability to differentiate these nuances of language is difficult to program into machines.

explicit representations of how verbs may relate to their arguments within a case grammar.

Discourse processing straddles the boundary between issues for IE that are specific to NLP, and those that are related to the nature of the specific IE task itself. Discourse processing, or co-reference resolution, is an area receiving a great deal of attention within the IE R&D field, because accuracy is critically dependent on these issues.

CHAPTER 3

INFORMATION EXTRACTION

Below is an overview of information extraction that distinguishes it from other related, but distinctly different advanced processing technologies. It includes a system-level description of a general application of IE and how it might fit into decision-support processes.

In simple terms, IE means "making information from data" based upon a predetermined need or criteria. The definition of IE presented in this paper targets all forms of data: speech (to include recordings, telephone conversations, and broadcast journalism); printed media (including archival materials); photographic and video media (still imagery and full motion); electronically transmitted broadcasts; digitally stored resources accessible through the Internet; terrain data; and data stored on local computing systems (databases). IE is an underlying technology that provides a single, albeit sophisticated, function that is not useful until it is designed into a system that performs a broader task.

Much of the focus within the literature addressing the subject of "Information Extraction" singles out the processing of text. However, potential information exists in any data form. Incorporating the term "Information Extraction" to mean only concerns for textual facts too narrowly construes the nature of the entire information processing function. Yet obviously, many improvements must be made to the basic underlying technology for speech and image processing before either category of sensory data can be applied at a level equivalent to that now being demonstrated for text.[26]

IE technology belongs within the branch of artificial intelligence science that studies expert systems theory. Expert systems attempt to mimic the way humans perform a given task. Specific types of information are gleaned through IE, based on the definitions established by organizational focus and/or by the expert-analyst.[27] The following describes how IE operates:

> One way to think of information extraction is in terms of database construction. An IE system attempts to convert unstructured text documents into codified database entries. Database entries might be drawn from a set of fixed values, or they can be actual substrings [words, phrases, letter groups] pulled from the original source text. For each IE task, a virtual database has to be designed which specifies the number of database entries that can be filled, and the data specifications for each of those fillers. For example, in an IE task on terrorism, we might define a priori [before hand] 24 possible terrorist activities. Then the slot for the terrorist event type must be filled with one of these 24 possible descriptors. But the slots that identify victims and perpetrators need to be open-ended. Those slots are best filled with strings from the source text.[28]

Extraction, Retrieval, Translation: Advanced Information Processes

Information extraction is a concept related to, but distinct from information retrieval. Information retrieval (IR) technology focuses on finding documents which contain topics related to the interests of the user generating the search. This is what a research student does when he or she goes to the library to find references

[26] Except when specifically stated within this paper, I will refer to IE as applying to all forms of data. Progress in speech recognition technology is much further along than that for imagery. Once speech-to-text becomes more reliable, only that conversion needs to occur before extraction tools can be applied.

[27] - This definition is somewhat wider than that applied by the government-funded research community (information extraction as only related to analysis of unrestricted, ASCII translated, text). Expert-analysts are key persons involved with a decision making process who review situationally relevant information and conclude specific values dependent upon its quality. These analysts could be employed as "experts", who might contribute to the establishment of an expert system which would automate an intelligent process. Examples of military positions which may be considered expert-analysts would be the intelligence officer (G/J-2), a program manager (PM), a commander (CDR), an operations officer (G/J-3), a logistician (G/J-4), a medical researcher, judge advocate officer, etc.

[28] Wendy Lehnert, Professor of Computer Science at the University of Massachusetts in a document titled "Information Extraction", available through the World-Wide Web, an Internet resource at the U. Mass. home page address, http://www.cs.umass.edu

for his or her topic. The student types an automated query on the library's computer system, based on title, author or subject, data; if materials are available that match the student's query criteria (title, author, subject) a list is displayed indicating where the documents are located.[29] Information retrieval can be thought of either as a preparatory step in a text IE application, or as a separate process whose output is the input for the IE system.

In contrast to IR, an IE system navigates through text documents looking for specific kinds, or classes of information. When data of interest is discovered it is copied to the IE system's database using a templates as placeholder structures. Succinctly, IR identifies where specific information is located, while extraction identifies the relevant data that is in a specific document.

Machine Translation (MT) is concerned with the automated interpretation and translation of foreign text into a language familiar to its user. While an MT system can convert an entire document to a known language its functions are only applicable to text, and not speech. Current speech recognition capability is not mature enough to integrate with MT to form a reliable, real-time MT capability for voice input.

There is a tight coupling between natural language IE, IR, and MT since the underpinnings of each process are similar. All retrieval systems use natural language processing techniques to identify documents. IE uses natural language methods to identify relevant terms and entities contained within narratives.[30] MT relies on natural language processing to interpret foreign language structure and convert it into a pragmatic representation in a desired human language. The

[29] Within the literature, information retrieval is sometimes associated with the term "data spotting", sense the focus of retrieval is to find documents that may contain the words in the subject, or query, specified by the user. Similar in idea to a "search" or "find" function in a word-processing software package.

[30] The terms narrative, free-text, unformatted text, raw text, are synonymous within this paper.

purposes of IE and IR are similar in that both IE and IR filter large datasets into a reduced, more consumable collection, while MT has the unique purpose of language translation.[31]

It must be clarified that these systems do not attempt to interpret the entire text in all input documents, but rather focus on those portions of each document that contain relevant information.[32] Relevance is established by predetermined domain guidelines, or rules, that specify exactly what types of information the automated system is expected to extract. Machines that have the ability to fully understand natural language in any form, may be a futuristic possibility for the human language technology research area.

In summary, information extraction technology is used to pull information from a mass of data. It is different than other natural language processing technologies, because only IE can interpret an information value from raw data.

The Information Extraction Process Model

The model for IE consists of several phases that logically parallel those found in already existing information systems: Data is collected from a source, translated into a common format, placed in temporary storage awaiting processing, processed by the extraction system, validated for quality and accuracy, and distributed to an application or user. This sequential process requires completion of each preceding step before advancement to the next one. Specific stages of the IE process model follow.

[31] Mary Ellen Okurowski. "Information Extraction Overview", Proceedings - TIPSTER Text Program (Phase I) San Francisco: Morgan Kaufmann, Sept. 1993, pg. 117. This reference contains a summarized definition of information extraction and the distinction among the three related technologies.

[32] Text understanding, the recovery of all information contained explicitly or implied within a text document, is extremely difficult. The scope of text understanding presents a number of problems that have not been adequately solved. These problems entail the full complexity of natural languages, and in particular the complete syntactic and semantic analysis of sentences.

Raw Data The raw data examined by IE consists of communication entities that I have collectively referred to as documents. Documents include both products such as magazine articles, newspapers, news services, broadcasts, and also proprietary communications such as memoranda, formatted messages, electronic mail, telephone conversations, photographs, terrain data products, video-teleconferences, and records.[33]

Collection. Identification is the first step in collecting data for IE. The proliferation of electronic information sources and formats dictates the need for additional scrutiny in selecting the optimal sources for consideration. Since a common character format is required by the hypothetical text extraction system, it becomes advantageous to use sources that communicate using the American Code Information Interchange (ASCII) character set. A similar issue exists for imagery formats with several graphic formats such as TIF, GIF, PCX in use.[34]

Acquisition generally involves establishing an electronic interface to a source and retrieving documents. These interfaces may include a document imaging equipment, a connection to information servers, a broadcast channel, a satellite down link, or an existing database. The collected data is brought into the domain of the IE system and either placed in temporary storage or processed immediately.

Translation. If the data source does not communicate in the language understood by the IE system (e.g. either in a natural language form-such as English, or the

[33] I chose to present a listing of the kinds of materials that may be used as raw inputs to IE systems. This was done to establish the breadth of scope that I believe IE to embrace. Webster's II New Riverside University Dictionary (Boston: Houghton Mifflin) 1988, defines a document as "something serving as proof or evidence,...". It is this broader definition that I apply in using that term in this paper.

[34] A proposed national imagery standard undergoing approval process.

machine text language ASCII), then a translation step is engaged.[35] In the case of sources that use speech recognition, voice-to-text recognition processing is employed. In the case of imagery, sources would be converted to the selected common form. Printed documents, not accessible in electronic media, would be scanned in as images, and have optical character recognition (OCR) systems convert the image to ASCII text.

Extraction. After the previous steps have been verified, data is processed into a meaningful form. The product of IE is called a template. To summarize, the stages within the natural language extraction process can be assumed to comprise:

1. A relevance filtering step that discards irrelevant information and flags relevant sentences.

2. Sentence level processing that extracts information from sentence fragments (phrases, clauses, sentences).

3. Discourse processing that establishes co-reference and merges events that apply to the same object.

4. Template generation that maps events into templates.[36]

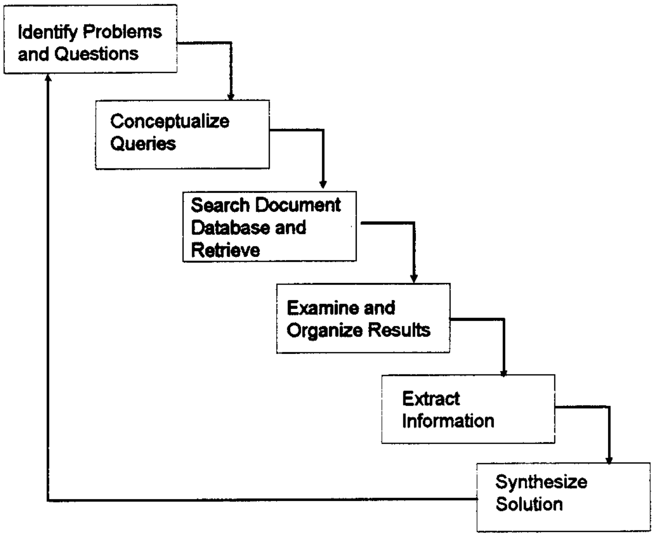
Imagery and terrain data (which is often referred to inclusively with imagery) are processed on a conceptual level, with information extracted to meet the terms of the concept employed. The extraction of information from full-motion video is a new and challenging niche in the area of information processing research and development. A discussion of these imagery and video considerations is presented later as well.

[35] If language conversion is required, a machine translation system could be applied at this stage.

[36] Lynette Hirschman, "An Adjunct Test for Discourse Processing in MUC-4", appearing in Proceedings - Fourth Message Understanding Conference (MUC-4). San Mateo, CA: Morgan Kaufmann pg. 67.

A Document Based Information Model Incorporating IE

There has already been significant development in the area of document processing and document based IE. A representative diagram is included below and an outline of the concepts it represents is provided for a more holistic appreciation of IE.[37] This is a generalized interpretation using the diagrammatic representation of a single developmental system that uses text-based IE. Nonetheless, this particular diagram is instructive in demonstrating how IE fits a potential information application, and serves to amplify IE features described earlier.


Identify Problems and Questions
Conceptualize Queries
Search Document Database and Retrieve
Examine and Organize Results
Extract Information
Synthesize Solution

Document Based Information Model

Figure 2

The model begins with the identification of the problem and related questions. While information extraction technology holds the potential to automate

[37] Original form of this diagram taken from the National Air Intelligence Center's (NAIC) Document Content Analysis and Retrieval System (DCARS). DCARS literature provided by Calspan Advanced Technology Center, Box 400, Buffalo, NY 14225. I have modified the diagram and provided my own interpretation.

data processing steps, proficiency in user problem solving still remains critical to maximize IE utility.

The model's second step concerns the conceptualization of queries using natural language interpretation tools, Boolean techniques and manipulation of database fields.[38] Concept based queries of large datasets require sophistication in both the target application information system being manipulated. This is an area in which IE R&D may provide significant insight, by developing an understanding of natural language processing that can be applied to the formulation of complex queries.[39]

The information retrieval segment is the third step and identifies all documents in a dataset containing references to pre-determined key words and phrases. This step may be thought of as a "first-cut," or filtering of a dataset which determines the quantity of subject-specific material that is even remotely concerned with the topic of interest. Once documents are indexed for examination, tools related to IR may be used to visualize and spatially arrange the data for better understanding and further processing.[40]

The objective in utilizing this model of information processes is to exploit the capability to automatically extract data from a dataset of unknown quality.[41] The extraction step reviews the relevant documents, as determined by the IR step, and

[38] A Boolean formed query is of the form <this> AND <that> OR <that>.

[39] Conquest Software Inc. "NetQuest: Natural Language Access to Distributed, Heterogeneous Information Sources": Natural language query capability, if accurate, can transcend much of the overhead required in working with large datasets. If this kind of data manipulation can be achieved, normal thought and speech patterns could direct the extraction of information, through voice recognition or other advanced machine-interface techniques.

[40] Visualization tools provide a multidimensional, graphical representation of the dataset that results from the initial IR query. This is a technical niche of its own right.

[41] "Unknown quality" means that there is an unknown probability for the occurrence of documents containing information of interest to the analyst.

brings in information relevant to the problem identified in the first step.

The synthesizing of solutions, the last step, is ultimately the focus for all of the intermittent information processes within the model. The model is represented here as a continuous decision support/command and control process, that requires data until all decisions related to a situation (a series of problems and questions to be solved) can be rendered and actions be completed.

Natural Language Extraction

There is a strong linkage between the previous discussion of NLP and the description of an IE system that follows. Jerry Hobbs, of SRI International's Artificial Research Center, describes a generic information extraction system which employs natural language processing methods. I use his description as the basis for my IE system diagram and functional outline for automated information extraction.

Hobbs defines an IE system as a "cascade of transducers or modules that at each step add structure and often lose information, hopefully irrelevant, by applying rules that are acquired manually or automatically."[42] The outline of a generic IE system consists of a set of functional modules which reduce the amount of data and evolve it into useful information. The stages of Hobbs' generic model for IE can be grouped into classes based on each general function. The four general functions and the corresponding stages from Hobbs' model are: filtering (text zoner, preprocessor, filter); parsing (pre-parser, parser, fragment combination); content understanding (semantic interpretation, lexical disambiguation, and coreference resolution); and lastly, production (template generation).

[42] Jerry H. Hobbs, The Generic Information Extraction System, Proceedings of the Fifth Message Understanding Conference (MUC-5) August 1993. (San Francisco: Morgan Kaufmann Publishers, Inc, 1993) p. 87.

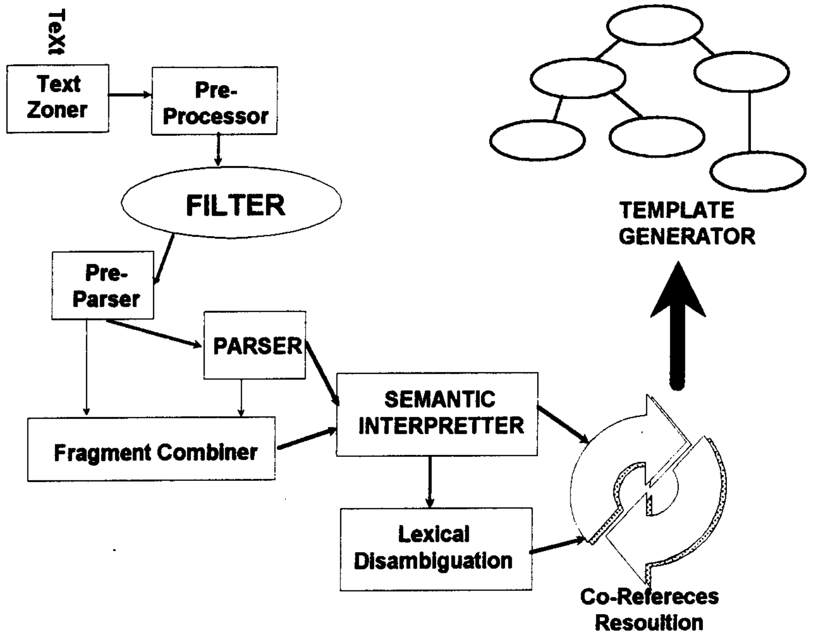


Figure 2 - Generic Information Extraction Process Model

Filtering. In a general sense, filtering, is the process of discerning relevant from irrelevant text. The goal of filtering tasks is to make the remaining IE processes easier by eliminating extraneous text. One context for filtering occurs at a document level: if a document contains relevant data or it is evaluated by the IE system, otherwise it is discarded (not evaluated). Another context for filtering is within a document of interest. At this level, only text that has characteristics that may be of interest are passed to later stages, the remainder is discarded.[43] The components of Hobbs' model that perform a filtering function are described below:

The **text zoner** module may separate formatted regions (usually the text header) from the unformatted regions of text (the main body), or sub-topic areas and their corresponding texts from the main text.

[43] David D. Lewis and Richard M. Tong. "Text Filtering in MUC-3 and MUC-4", Proceedings - Fourth Message Understanding Conference (MUC-4). San Mateo: Morgan Kaufmann, pg 51. A related discussion of filtering, not specific to the generic IE model, is presented in this work.

The **preprocessor** module takes groups of characters and filters based on the parts of speech they represent (nouns, verbs, adjectives etc.) Dictionary lookup and morphological analysis (the form a root stem takes within the content of a text)of words occurs within this module to help distinguish parts of speech.

The **filter** sifts out relevant (pre-defined as part of the IE architectural design) from irrelevant sentences by recognizing keywords, word-patterns or n-grams (character patterns of "n" characters in length).

Parsing. The parsing function, which was described in earlier section on NLP background. Parsing identifies relationships between language terms using a set of rules called a grammar. The modules from Hobbs' generic IE model that participate in parsing functions are described:

The **preparser** module identifies common small-scale text structures such as clauses, noun, verb and prepositional phrases.

The **parsing** module a produces completed parse tree.

Parse tree fragments result when the parsing program cannot completely parse a sentence into a grammatically correct structure. The **fragment combination** module attempts to resolve complex parse tree fragments into data, or upon failing to do so, discards them.

Content Understanding. This general function renders meaning from text segments. Much of the work in this function concerns the resolution of terms and decisions on the meaning of terms that the system will provide to the template production stage. As suggested in the section on natural language processing, this function should be influenced by the application that uses the IE product to achieve practical results.

The **semantic interpretation** module translates the parse tree into a logical relationship of understood language terms (i.e. predicate-argument form).[44]

The **lexical disambiguation** module translates recognized, but ambiguous, language terms into unambiguous predicates. Lexical disambiguation "usually happens by constraining the interpretation by the context in which an ambiguous word occurs, perhaps together with the apriori probabilities of each of the word senses."[45]

The **coreference resolution** module processes the semantic structures (after lexical disambiguation) in the same manner as described for discourse processing in the section on NLP. A very simple example of this module might be the merger structures containing "the American secretary of defense" with instances within other semantic structures containing "William Perry."

Production. **Template generation** is the fundamental objective in the IE process. This module uses the structures created by the previous modules and produces a template in the object form required by the consumer of the information.

The reference citing Hobbs' paper on the generic IE model elaborates on the descriptions of the individual modules in his hypothetical system. I have attempted to capture the most important aspects of the IE process and to briefly outline a system level architecture that is consumable, without unnecessary detail.

[44] A predicate is a noun-verb grouping such as "she fights" or "the tank fired".

[45] Hobbs, pg. 90.

CHAPTER 4

NATURAL LANGUAGE PROCESSING RESEARCH & DEVELOPMENT

The Advanced Information Processing and Analysis Steering Group (AIPASG) is a body composed of representatives from the federal intelligence and scientific community. The steering committee provides oversight to government sponsored R&D designed to improve information processing capabilities that supports national security interests. AIPASG conducts an annual symposium which communicates the status of efforts within this particular area of research and provides an opportunity for developers to demonstrate their systems. The symposium hosts representatives from the national intelligence agencies, interested commercial developers, military organizations, federal agencies, and the Advanced Research Project Agency (ARPA).

TIPSTER (not an acronym) is an ARPA funded program dedicated to helping government intelligence activities detect and extract information from free-text documents. These documents include such diverse sources as books, newspapers, messages, and word processor created files, etc. The vision for the core technology under development in the TIPSTER program is the realization of a completely automated "intelligent assistant" capable of supplying the intelligence analyst with all necessary information organized to help with the specific analytical task at hand.

TIPSTER objectives have been divided into two phases. Phase I, completed in 1993, focused on methods for routing, retrieving, and extracting information from free text. Phase II, (completing its first year in March 95), seeks to refine the previous research into useful, deployable technology. A TIPSTER software architecture, whose purpose is to allow detection and extraction technologies to easily be deployed together and work synergistically, is scheduled for completion this month (May 95).

Background

The TIPSTER program was conceived after the second Message Understanding Conference (MUCK-II) in June of 1989.[46] The promising results of MUCK-II, coupled with the increased appreciation for the need to automatically handle large volumes of text, led to the formulation of a text-processing technology development plan. Advanced Research Projects Agency (ARPA, formerly DARPA), the NSA, and the CIA jointly funded TIPSTER hoping to achieve substantial benefits for the government's need for sorting and analyzing immense volumes of text. TIPSTER undertook separate focuses for detection, (information retrieval and routing; correlated through the Text Retrieval Conferences-TREC), and on extraction (the forum for this focus is the MUC series). These two capabilities are considered to underlay all other information handling tasks. The sponsorship program is generally arranged so that ARPA contracts for the development of a standard evaluation methodology and a common test corpura, (bodies of documents) as raw data, to be used in the evaluation of systems presented at either of the two conferences.

Message Understanding Conferences

The Message Understanding Conference (MUC) series sessions have highlighted the ongoing development of systems capable of analyzing textual presentation of data, and the definition of a performance evaluation methodology.

[46] The first and second MUCs bore the acronym "MUCK"-I & II, vice MUC-3, etc., for the subsequent conference sessions. MUC-1 and MUC-2 were organized by NRaD and NIST before sponsorship of these programs became part of TIPSTER.

The Naval Command, Control, and Ocean Surveillance Center (formerly the Naval Ocean Systems Center), has been a principal participant in this research area, hosting the first three conference sessions. Each session has had a slightly different focus, the composite effect being a maturation of the IE field of research. A synopsis of each session follows:

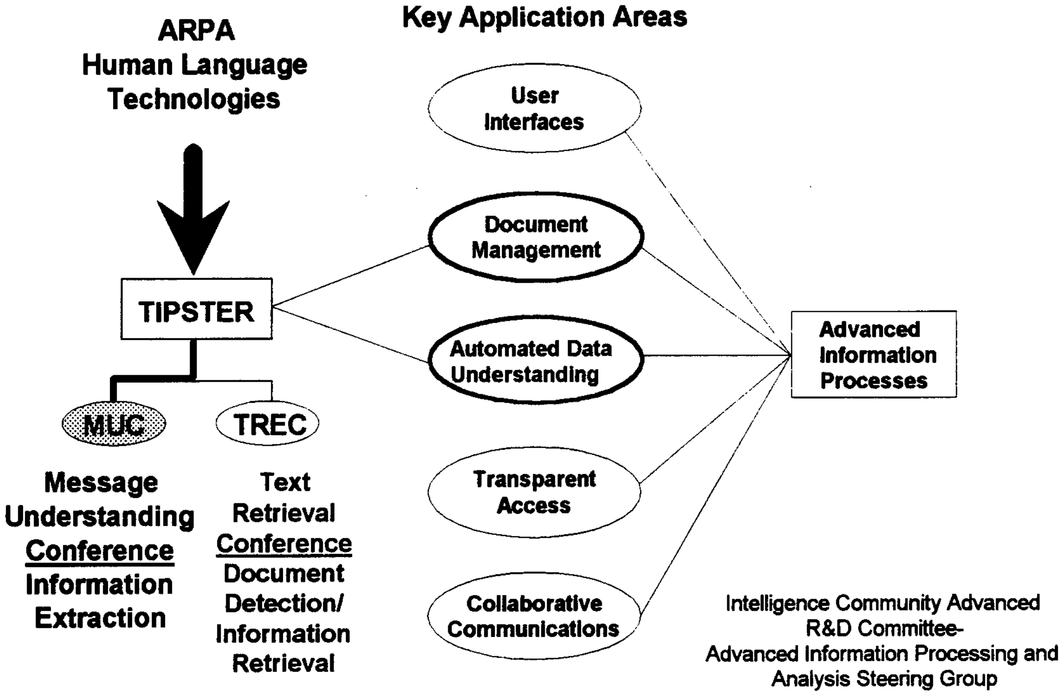


Figure 3 - Information Extraction R&D Model

The First Message Understanding Conference (MUCK-1, 1987) used a small body of Naval message texts as the data for testing applications presented by natural language processing (NLP) systems developers. This session demonstrated the potential for machines to extract significant amounts of text from semi-formatted sources.

Second Message Understanding Conference (MUCK-II, 1989) shifted the focus to a quantitative evaluation of NLP systems, using an expanded body of Naval text messages as the data source. MUCK-II intended to test the text analysis capabilities of NLP systems that could analyze short (paragraph-length) segments taken from military messages.[47] MUCK-II proved that some existing systems could reasonably perform data extraction from poorly formed text paragraphs, albeit in a relatively narrow domain.

Third Message Understanding Conference (MUC-3, 1991) efforts addressed information extraction from continuous (unformated) text. MUC-3 systems attempted to analyze articles distributed by the U.S. government's Foreign Broadcast Information Service (FBIS). The MUC-3 evaluation simulated potential real-life scenarios (such as the recent sarin nerve agent incident in Tokyo) in which intelligence agency terrorism specialists require automated assistance to process a high density of open-source data from news services.[48] Experiences with MUC-3 systems suggested that applications for computer-assisted extraction from unconstrained text had the potential to improve human analytical productivity by a factor of five.

Fourth Message Understanding Conference (MUC-4, 1992) Problems in discourse reference resolution and inferencing were identified. As an example of discourse resolution, the inability to reliably determine whether a description found in one part of the text refers or does not refer to something previously described inhibits both recall and precision because it could result in either missed or spurious information; the inability to pick up subtle cues to relevant information places a

[47] Sundheim, Beth Second Message Understanding Conference (MUCK-II) Report Naval Ocean Systems Center Technical Report 1328, September 1989 (San Diego: 1989), 1. Both MUCK-I and Muck-II centered around the analysis of telegraphic-style texts relating to confrontations between friendly and hostile naval forces.

[48] Beth M. Sundheim, and Wendy G. Lehnert. "A Performance Evaluation of Text-Analysis Technologies," AI Magazine, Fall 1991, 81-94.

limitation on recall because it results in missed information. [49]

Fifth Message Understanding Conference (MUC-5, 1993) subjected participating IE systems to increased levels of complexity over that found in MUC-4. Text articles from two topic domains (versus the terrorism only domain in MUCs-3 & 4), and in two languages (Japanese and English) formed the test body.

Steady progress has been reported throughout the TIPSTER and MUC series. There appears to be much promise for improving the ability of machines to read and interpret data from free-text. Machines are capable of processing several times as much text as compared to humans but are not as accurate. ARPA's assessment is that the most rapid integration of machine processing of information may be in "human-in-the-loop applications", where an analyst or researcher is supported by one, or several, IE systems. In these applications, the analyst's tasks may be oriented more on insuring machine accuracy rather than extracting data.

Machine Performance Compared to Human Performance

The MUC-5 included a disciplined study of human performance compared to machines.[50] Humans and machines used text articles from magazines and newspapers that were part of MUC-5 test data set. This data consisted of articles, that were concerned with joint Japanese and U.S. business ventures and microelectronic fabrication technology. Below is a "thumbnail" comparison reflecting the general results of testing is presented below. The information in the table indicates that the machines are much faster at processing a document, but are

[49] Defense Advanced Research Projects Agency Fourth Message Understanding Conference (MUC-4). (San Mateo, CA: 1992).

[50] Craig A. Will. "Comparing Human and Machine Performance for Natural Language Extraction: Results for English Microelectronics from the MUC-5 Evaluation", appearing in Proceedings of the Fifth Message Understanding Conference (MUC-5) pg . 53.

only about half as accurate as humans.[51]

Comparison of Performance Averages (Using MUC-5 Test Data)[52]

Measure System	Recall	Precision	Estimated Time/Document	Amount of Accurate Info
Machines	53%	57%	3 Minutes	31%
Humans	79%	82%	30 Minutes	65%

Performance Metrics

In order to design research projects and monitor developmental systems, a set of metrics has been formalized. The measure of a system's ability to extract information from a known (test) dataset is called "recall." It is a measure of the amount of data extracted from the dataset, compared to the known, or expected amount of information to be extracted from the dataset. The measure of an IE system's accuracy in extracting information from a known dataset is called "precision." Precision is determined by the amount of data correctly extracted as a ratio of the amount of information actually in the dataset. Both recall and precision are expressed as percentages. A clarifying example is provided by Professor Wendy Lehnert, of the University of Massachusetts:

> To better understand recall and precision, suppose you took a True and False test and got 1 out of 4 questions correct. That would be 25% recall. Your precision would depend on how many questions you actually tried to answer. If you answered only 1 and left 3 blank, then you would have 100% precision. But if you answered 2, then you would have 50% precision.

[51] The references note that many problematic factors could influence the reliability of these results. However, the purpose of presenting this information is to merely give the reader an estimate of how this particular technology scales with respect to present human benchmarks.

[52] Wendy Lehnert, Professor of Computer Science at the University off Massachusetts, "Information Extraction", available through the WorldWide Web at the UMass home page address: http://www.cs.umass.edu. Information in the table is presented here.

Precision and recall are widely used metrics that can be applied to many individual data types (text, imagery, databases, speech). These measures are however difficult to apply to iterative or multiple searches and do not assess query formulation (designing the question in the right way), result merging (collecting all results into a group), or information presentation (how is the information displayed to the user). Additionally, precision and recall are not applicable to structured data since modern query languages ensure that both precision and recall are 100% assuming that the right query is asked of the right database.

Issues in Evaluation Metrics

An analysis of evaluation techniques for integrated data access methods, such as information extraction systems, was presented by Dr. Len Seligmann and Dr. Marcia Kerchner of Mitre Corporation, at the Advanced Information Processing and Analysis Symposium held in March 1995.[53] The research question for their set of issues was "How do we assess the extent to which different approaches actually help put useful information into the hands of users?". The project sought to determine whether or not comparative evaluations of alternate approaches could be done, and if currently used measures sufficiently characterize the value of integrated data analysis (IDA) technologies. Doctors Seligmann and Kerchner used four categories of measures in their evaluation:

Relevance-based measures that are used to evaluate a technique on its ability to extract a given information item in response to an information need. This measure is highly subjective and context-dependent with respect to a given

[53] Sponsored by the Advanced Information Processing and Analysis Steering Group (AIPASG) of the U.S. government's intelligence community. Chair of the AIPASG is Ms. Anita I. Cohen; symposium chair as Dr. Russel R. Rose. The symposium was held 28-30 March 1995, at the Sheraton Premiere Hotel located in Tysons Corner, VA.

application. Relevance measures are of specific interest to current military application because of the significance placed on relevant information for the commander. The filtering and fusing of the right information (situationally relevant information) is deemed critical for operational success in the information age.[54]

Utility metrics attempt to evaluate the value-added by a given information technique. The components of utility are informativeness and overall value of information. The informativeness of a search trail measures the extent to which the search corresponds to some ideal answer trail (if the best possible answer existed within some accessible dataset, how close would have the evaluated system been able to providing that answer).

The overall value of the information provided represents a subjective evaluation of how much the information provided by a specific technique or system contributed to a solution to the information requirement.

Measures based on Cost of Searching, or the time a user must spend using on more techniques and systems, cover all aspects of data retrieval and information extraction to include query formulation, fusing of results, and presentation of information. This metric can alternatively be converted to a monetary cost analysis.

User Satisfaction is a subjective measure that may also be employed to gauge the effectiveness of information services.

Considerations for Developing Relevant Evaluation Criteria for IE

There are many opportunities to apply IE technology on actual datasets in order to study IE's potential to assist in real world applications. Events from the past may be identified, and the archived, relevant dataset that was available during the event time frame may be re-created. IE systems can be run against archived

[54] U.S. Army Dept. Information Operations (Coordinating Draft), FM 100-6 , n.p. July 1994, various areas.

datasets to produce templates that might have occurred during the actual event period. Results from these tests can determine the likelihood with which IE systems might have significantly contributed to decisions made during these events.

Introspective revelations pertaining to command and control or mission planning processes may be garnered as a by-product from performance testing that uses actual operational data.[55] For example, IE could be applied to data available during the invasions Grenada and Panama, or prior to the assassination of COL Nick Rowe in the Philippines.

[55] This is akin to how the military does operational testing as part of the system acquisition process.

CHAPTER 5

SUGGESTED MILITARY APPLICATION AREAS

Specific military applications for information extraction technology other than those for intelligence have not been articulated within the IE research community. This may be of distant concern, since the maturity level of IE technology cannot yet field effective operational systems. However, rapid progress experienced in clinical IE research promises to deliver automated capabilities that can be exploited to leverage military resources.

Results described in IE research literature suggest economies that might reduce administrative workload or improve effectiveness in existing military systems. However, researchers are not aware of the range of military problems to which IE technology could apply. Below are several specific applications from a number of functional disciplines that could integrate IE technology. Each application is described at a general level to highlight the potential for IE rather than to analyze specific military problems in detail.[56]

Human Resources Management

A contemporary, non-military example for IE employment is the human resource administration functions in electronic job banks.[57] Thousands of resumes

[56] Machine translation capability may be integrated within the scope of these applications to convert foreign language sources into English.

[57] Joe West, "How to Get a Computer to Read Your Resume", Army Times, 6 Feb 1995 pg R6.

submitted to job banks are processed by automated systems. Resumes submitted to a job bank may not normally follow a specific format. Even with semi-structured documents, provided in some forms, free-text must be reviewed to determine relevance. If a printed resume is sent to the job bank,. it is read by a scanner, stored, then converted to a standard character representation such as the American Standard Code Information Interchange set (ASCII). An IE tool is then applied to extract the predetermined information from applicants meeting the principal screening criteria. The IE tool may be applied directly if the resume is sent electronically, (already in ASCII form). In this manner, IE assists the human resource/personnel administrator in selecting best-suited applicants for interviews or for determining additional stages in the hiring process.

Variants of the human resources IE applications, such as the job placement example above, may be useful in military recruiting, career management, personnel surety (access to classified data) and various other administration tasks. Military job descriptions often are not good indicators of a service member's experience. Increased reliance on the narrative fields in personnel records would convey a more detailed assessment of an individual's knowledge and performance characteristics. To more effectively serve the human resource requirements of the services, IE tools could be integrated into a document management program that classifies, and extracts such useful personnel information. Equipped with IE tools, military organizations, especially those with non-standard and technical requirements (for example, foreign language, particular technical systems experience, or combat skills,) might be staffed more appropriately.

Army Promotion Boards and Officer Fitness Reports. The narrative portions of fitness reports have become more important as trends in personnel rating practices in the field have rendered ordinal discriminators inconsequential. For example, a recent Army Personnel Command (PERSCOM) report mentioned that most field grade officers are rated in the "top block" of the composite ranking section of the officer evaluation report. This ordinal evaluation technique was once used as a gauge for thumbnail sketches of an officer's past performance.[58] Inflated evaluation practices have forced promotion boards now to pay more attention to the narrative evaluation sections and spend more time on each officer's file. IE techniques could be employed to extract important free-text to facilitate during promotion screening boards.

Medical Services Support

Intelligent access to patients' medical records can be applied to extract diagnoses, symptoms, physical findings, test results, and therapeutic treatments. As the nature of battlefield medicine evolves with telecommunications systems, IE will be required to support "distant healing" concepts involving the telepresence of medical specialists. At least one case of this application has occurred involving a U.S. serviceman in Macedonia.

These systems can also be used to assist health care administrators in statistical research and quality assurance programs. A suggestion for the support of private sector medical insurance processing, whereby each patient encountered must be categorized for reimbursement purposes (easily transferred to the military's CHAMPUS or TRICARE programs for example) has been made in other IE literature.[59]

[58] The senior rater in an officer's chain indicates the rated officer's potential by symbolically placing him at the very top (top block) in the top third, upper or lower half, or at the bottom of his peer group based on the rater's experience with officer's he has been associated with at the rated officer's career level.

[59] Professor Wendy Lehnert of the University of Massachusetts discusses several non-military applications for IE in her paper "What is Information Extraction", available through the U Mass front page at http://www.cs.umass.edu.

Intelligence

IE technology, in its current state, is most applicable to military intelligence functions that process information. This capability is a direct transfer from the non-DoD intelligence community's adaptations for automated IE capability. I discuss the employment of IE for military intelligence applications using terrorism as a case in point. Terrorism response is interesting as an example because it encompasses a breadth of issues related to the roles, missions, and information support for the military. This specific IE capability is of particular interest to the military's special mission units whose operational focus includes anti-terrorism capability.

Terrorism. Terrorism response illustrates potential IE applications within the military intelligence function for four reasons. First, because anti-terrorist activities are functions of both non-military and military agencies and their interconnection at national level is critical to national security. Second, anti-terrorism efforts employ specific types of militarily related technical information, through both military and non-military information sources. This fact makes the scope of potentially relevant information very broad. Third, terrorism has been used as the topic area for test datasets involved in the TIPSTER program.[60] Lastly, recent large-scale, foreign and domestic terrorist activities raise new concerns for national security.

Over the last several years the natural language processing research and development community has used print and broadcast media containing reference to incidents of terrorism,. This data has been used in testing prospective IE systems to analyze their effectiveness. These prospective systems were evaluated for how well they identified the type of terrorist event, suspected perpetrators,

[60] The Advanced Research Project Agency's text information processing research and development program. TIPSTER is not an acronym.

fatality and injury figures, building damage, as well as the time and location of the event in question. Demonstrations of subject-specific extraction capability by developmental IE systems, have shown they may be of great benefit to multi-agency intelligence tasks.

The second aspect of IE in support of counter-terrorist intelligence is transaction monitoring. The purchase of the component chemicals for the nerve agent sarin, fissile materials, or large volumes of fertilizer with high nitrogen content, can be monitored using IE supported applications. These IE applications could monitor a wide range of free-text or semi-structured administrative data already available today, to include point-of-sale records, hazardous material transportation requests, truck/van rental agreements, publicly filed business contracts.

Extraction On-Demand: Intelligence Research

Automated IE techniques may be part of the answer to the problem of crisis-information requirements. Intelligence capabilities can use IE technology to "jump-start" the base of information accessible by military elements. Using model-based extraction, possibly in sequence with model-based retrieval, the power of computers would be harnessed to rapidly extract information pools.[61] These pools would be built to be of interest to a number of situational and functional concerns that evolve in the global environment. In this sense, as new subject areas develop (such as the break-up of the Soviet Union; Iranian nuclear capability, the military role in tribal violence in central Africa), automated and "intelligent" research could be undertaken using IE functionality.

[61] See chapter 7 for a discussion of model-based extraction and model based retrieval.

Crisis Intelligence... For the purposes of this paper a crisis is defined as an unanticipated or premature event that requires the rapid mobilization of significant resources. New information requirements could be designed into templates to extract information from existing text archives to support an unforeseen international problem. Because the general capabilities of IE tools, machines can outperform humans by a factor of 4 or 5,.[62] some aspects of crisis intelligence tasks can be compressed in time, freeing up human resources for more conceptually complex assignments.

Operational Role - Information Processing of Communications

Use of IE streamline the communications effort by extracting critical and relevant information from unformatted messages is certainly desirable. Government-funded research in IE began with analysis of tactical messages from naval exercises. This is likely the reason the ARPA-funded IE-technology development series was entitled the "Message Understanding Conference". This ARPA-sponsored research focused on battlespace related IE. The test databases of the first two Message Understanding Conferences consisted of simulated operational messages concerning hostile encounters between U.S. forces and a Soviet-like adversary.[63] The events included detecting, tracking, targeting, harassing and attacking activities. Messages were fed into IE systems to create databases for U.S. and adversarial activities. The success of this testing serves as a proof of concept for continued IE research, and also proves that IE has a role within battlespace C2 and operations.

[62] Craig A. Will. "Comparing Human and Machine Performance for Natural Language Information Extraction", Proceedings - Fifth Message Understanding Conference (MUC-5). San Mateo, CA: 1993, p. 53.

[63] Beth Sundheim, "Second Message Understanding Conference (MUCK-II) Report", Technical Report 1328, September 1989, San Diego: Naval Ocean Systems Center, pg. 10.

IE activities may be extended to telephone conversations, facsimiles, orations, briefings, studies, reports and other administrative communication. The key benefit of this IE application for the military is that even formatted messages contain narrative, or free-text sections that contain relevant information. When thousands of messages are generated during intensive operational settings managing and controlling the information flow is an extremely difficult, yet vital task. IE technology can assist by reading messages as they arrive, screening for the commander's information requirements (filling templates), identifying non-addressed units that may need to have the message forwarded, and cataloging the arrival and contents of the message for further processing by other humans or machines. Given the early experience within the MUC/TIPSTER program, this is a real capability that has been demonstrated in a developmental form. The incorporation of some form of IE capability within future communications architectures is anticipated in the near future.

Operational Research and Systems Analysis (ORSA)

U.S. military doctrine evolves through analysis of operations and the development of conceptual models of the adversary. The ORSA field requires the review of textual, and imagery data, a process that is extremely tedious but one that can be improved with IE technology.

Friendly-forces operations journals that contain spot reports, narrative recounts, and after-action reports submitted by subordinate units could serve as input to an IE system. Captured enemy documents or interviews of prisoners, could also be screened by an IE application to filter and collate enemy operational information. Subsequently, this information could be used by analysts in improving our interpretation of adversary's order-of-battle (OOB). In particular, reports of SCUD launches, reconnaissance data concerning suspected or prepared launch sites, and other theater missile deployment information could be automatically entered into databases, which would then be available to expand knowledge of enemy maneuver characteristics. This information also could be used in wargaming, to bring the OOB up to date, and for assessing opposing force weapons employment.

Modelling and Simulation (MODSIM)

IE of imagery information can to be applied to military-related modelling and simulation. Concept extraction techniques employed to find specific images can later be manipulated by modelling applications or simulation systems. Object linking of image concepts could result in automatic updates to the images presented in simulation activities (training, wargaming, rehearsals). This may be especially important in preparation (rehearsal) of the execution sequence for strike operations or raids, because target information may not be available early in the execution cycle. Once new imagery is acquired from national-level databases, standing critical interest areas may be automatically extracted and rendered to the MODSIM activity.

The principles of concept extraction could be used as a basis to direct data collection. In this capacity, an initial fact may help to substantiate a hypothesis being used to build a model or simulation for a military concern. As the model is updated through recursive application of concept extraction,[64] gaps in information, that may otherwise not have been apparent would be identified and collection plans modified to meet requirements.

[64] "Recursive" use of concept extraction techniques, means that concept extraction is used repeatedly to focus the level of abstraction from higher to lower order. Generally, models are created from a "top down" approach, and the details are worked out sequentially, once the high-order representation of reality (the problem) is validated.

Systems Acquisition, Research and Development

Systems have been designed to monitor technical articles describing microelectronic chip fabrication to capture information processing technologies. People in the Defense R&D, and systems acquisition career fields have a particularly acute need to be on the cutting edge of technology advances, because of the vast breadth of responsibilities the military program manager assumes. Yet the ability to submit a query on a subject area and receiving information from all known and available sources is non-existent. Information of interest to a program manager, such as contractor financial transactions and their status, the results of a contractor's work for other government and non-government projects relying on similar technical concerns, could be provided quicker through IE. Templating techniques can tailor the construction of information resources available to the program manager to make information analysis more timely and effective.

Transcending Politics and Culture

Cultural or political biases may cause readers to overlook relevant information in a document. For example, military personnel with a negative bias toward public media may completely disregard useful information from these sources . One use for IE within the military environment may be to make information processes more objective and less political by automating some processing tasks.

The value of a piece of information passed internally may be degraded by the organizational, or positional status of the producer. The personality, culture, or institutional background of the author (individual or organization) of a report may create questions of incredulity. A hierarchy of relevance biased on organizational, or even national politics is historically disastrous as our Vietnam experience documents.

Professor Michael Handel stresses the negative impact of the "politicization of information"[65] in his *Intelligence Policy and War* course taught at the Naval War College. When the ego of a commander is strong or once the commander has determined to undertake a particular course of action, intelligence staff officers are unlikely to communicate "bad news" effectively. Hence, reliable information that is contrary to the spirit of the commander's will, may be discounted, not reported, or massaged into acceptability all in the spirit of "teamwork". A technology such as IE may successfully automate the filtering, processing and analyzing tasks that are most susceptible to being corrupted through politicization. IE may have the potential to objectively cull information from the great mass of text, video, and spoken documents, without the social impediments that prevent accurate analysis of data.

Final Thoughts on Military Applications for IE

This paper suggests where IE technology may be employed to help leverage shrinking military resources. Although the discussion of military applications for IE is not conclusive, and IE technology may not effectively mature for some time, the applicability to certain military systems may be impractical due to sophistication in the military environment. Nonetheless, the employment schemes which are proposed here may stimulate IE research and development, in a direction toward specific military applications.

[65] I will use "politicization" to mean the collective effect of social, cultural and organizational biases that degrade the integrity of an item of information.

CHAPTER 6

VULNERABILITIES AND THREATS POSED BY IE TECHNOLOGY

New IE technology paradoxically poses both an advantage and a threat to our national security. The global proliferation of IE capability raises legitimate concerns about the vulnerability of defense information architectures.

Vulnerabilities

General vulnerabilities may originate because IE capability convinces analysts and decision makers alike that the intelligence information extracted by machines is an improvement over previous manual systems. While automated IE may deliver more information from more sources in less time and at reduced costs, IE may not always provide greater reliability for the information product.

The Automation Paradox. Assuming that the number of sources relevant to any situation will increase, when a greater quantity of information is automatically extracted, there inevitably also will be an ensuing increase in the amount of verification required. Given this verification undoubtedly will be undertaken with greater reliance on human (manual) efforts, thereby increasing the requirement for human resources for these tasks above today's manning levels. Hence. although IE may increase the amount of accurate, relevant and reliable information available to decision makers and command and control processes, it also increases the level of resources required to accomplish this task.

Additional requirements for system maintenance and modernization may also significantly increase the total resource requirement for a given agency. A high level of understanding in the related technical fields will be necessary for the users (analysts) and systems specialists (technicians) to jointly determine whether the payoff of IE technology investment can be measured objectively.

Fancied Technologies. American infatuation with technology as a solution to complex problems may be the cause of an over-acceptance phenomena. IE technology is not yet mature; however, the nature of the research and development business dictates that only those systems demonstrating the most advanced capabilities are chosen for continued funding. Hence, developers often are likely to give their systems the benefit of the doubt when reporting on capabilities in quantitative terms. Decision makers may be sold systems that at time of delivery cannot produce anticipated services, thereby creating an information gap between needs and servicing capability. While this potential pitfall faces all other types of systems, IE systems are especially vulnerable because they are positioned near the crucible of the decision making process. Although promises from this technology are part of the IE R&D objective, they should be viewed in the same light as other technical solutions for defense problems. After all, the "desktop" PC revolution is behind us, and clearly, the PC's have won![66]

Threats

Threats Posed by a Promiscuous Information Source. American security has been vulnerable for quite some time, due to its "information promiscuity." The U.S.

[66] The military's experience with other information systems developments provides ample historical perspective regarding fielding and integration misconceptions.

Constitution guarantees freedom of the press, a right that has invoked controversy throughout the history of our nation. Undeniably, there are more information venues originating from within the political boundary of the United States then anywhere on the globe. Information is readily sought and, except for some of the highest-valued information, protected by government and industry, most information is easily obtained. Ironically, by identifying the critical sources of data, IE capability could be made to automatically acquire and update a large mass of information that might then be used to exploit U.S. national interests.

Vulnerabilities to our national security interests become exacerbated by the expanded consumption capacities attained through automated information understanding tools. The United States is already the world's most prolific producer of open-source information. The advent of the "info-bahn," or information super-highway, as a medium for cultures of communication, as well as data exchange, promises to increase the amount of information that may be obtained for the price of an Internet access fee.

The stories of computer "crackers,"[67] individuals who gain access and cause damage to controlled and sensitive areas of information located on computers linked to the Internet are legend. What IE throws into the fray is not a destructive or duplicitous technique to surreptitiously acquire or alter information, but more dangerously, the availability of a potentially powerful method of understanding or inferring American intentions and cultural vulnerabilities. This can be made possible by automatically processing a greater number of American sources. These include communiqués of the niche, fringe, and special interest segments of American society, which give an adversary a better understanding of how to influence the national will.

[67] Crackers gain illicit access to information systems and cause destruction. Hackers, on the other hand gain access but do not do damage.

Threats Posed by the Global Proliferation of Technology. Another IE security vulnerability is the threat that extraction systems developed here or abroad will fall into the hands of adversaries. A general aspect of advanced information processing technology is that systems are assembled from off-the-shelf hardware and software components, which makes proliferation extremely hard to control. Major General Kenneth Minahan, Air Force Chief of Staff for Intelligence, describes this vulnerability in economic terms, predicting that normal market conditions will create an excess supply of technology. The threat from the proliferation of advanced capabilities to potential adversaries becomes imminent as producers turn to any source of revenue they can find to stay in business.[68]

Further, the capital costs for automated IE tools will be relatively minimal; present demonstration platforms are assembled from off-the-shelf hardware components for less than fifty-thousand dollars. The real technical capability is delivered through software development, which is where the bulk of IE research and development resources are applied. Moreover, market factors will create opportunities for potential adversaries to easily acquire all the necessary components, and perhaps at a bargain price.

The vulnerabilities posed by free access to information that could expose many aspects of US national security, coupled with the availability of advanced processing systems, could provide a shrewd foreign leader an upper hand to manipulate US policy during crisis situations.

Mutually Assured Extraction. IE may be employed as a front-end filtering and database-building capability that makes it feasible to examine nearly every piece of information that travels across a common medium, extracting data for concepts of

[68] During a keynote address to the Advanced Information Processing and Analysis Symposium, held at Tysons Corner, VA during March 1995.

interest to the receiver. IE can therefore be used as a means in the decision support mechanisms employed by our potential adversaries, much as we intend to use the same tool against them.

Misinformation and Disinformation. A reliance on IE to provide relevant information to meet U.S. national security interests should be viewed symmetrically with acute attention given to an adversary's attempts to thwart our technological advantages.

Potential opponents might employ "disinformation" as a counter-IE technique. Automatically extracted information, whether from the open-source journalistic media, compromised publications of foreign governments or information produced by non-governmental organizations (NGOs) will need to undergo the same verification process currently undertaken for manually extracted information. The ease with which disinformation can be interjected into the broadcast and print media is a particular cause for concern.

Misinformation, or the normal analytical failure often occurring in media reports concerning complex situations, often coupled with the willingness of news services to communicate unsubstantiated observations as truth, often leads to the conveyance of misinformation.

Concerns about disinformation and misinformation degrades the reliability, and therefore the value, of information extracted from any source not having undergone a legitimate verification process.

IE and Information Warfare. Should an enemy obtain a sophisticated understanding of how U.S. IE systems perform, he might be able to falsely attenuate parts of our security and defense apparatus for his own purposes.

The following two scenarios involve the exploitation of IE capability to threaten U.S. interests. Both of the scenarios highlight the military operational principle of leverage, or economy of force, which is a general characteristic of all information warfare activities. The first example demonstrates how IE capability could be exploited by an internal adversary to national security to pursue its objectives using information warfare.

A future example of threats to civil order having likely military implications, can be drawn from the events surrounding the recent Oklahoma City bombing and the many "militia" organizations highlighted in the national media. Envision a situation where writings in militia publications, which are processed through IE templating, provide indication of the potential for an act of large-scale violence and destruction. The overworked analysts, who rely on IE to perform the bulk of "trivial" tasks,[69] on reviewing these results, can neither uphold or refute this evidence. This is because there may be no contra-indicatory evidence (even though there may be only incidental corroborating facts), and second, because of the politicization factors discussed in a previous section. As a result of the interpretation rendered through IE processing, law-enforcement agencies (possibly assisted by military units) then become involved in aggressive, and unpopular, administration of laws. This might be followed by an unspecified period of time when no violent acts are committed. Consequentially, public opinion would begin to weigh against the political entities directing these actions, as was the intention of the disinformation campaign.

The second scenario highlights the potential uses of IE for deception. Should an authoritarian nation seek to create artificial concerns for the United States, (a strategic level deception) it could begin to generate misinformation and disinformation to give the impression that they had developed a great deal of sophistication in nuclear weapons research, as an example case.

[69] Trivial in the sense that information extraction is one of the simpler tasks to master, and because it involves much tedium -reading countless documents to yield relatively small amounts of relevant information.

If a nation wished to create the perception that it had a weapon of mass-destruction (chemical, biological or nuclear) at its disposal, it could gain a library of valid and accurate research information by extracting technical knowledge from countless, open-source, technical publications, text-books, and conference proceedings. Information could then be translated into the national language to create seemingly indigenous publications (post-dating and deceptively archiving them) for release through the R&D community. To insure that the deception operation was not discovered, the information warfare apparatus of the rogue country could seed information into obscure or less well read publications (but assumed to be used as IE sources) as well as an occasional "first-tier" publication such as the New York Times.[70] This deception would be accomplished over a period of time to allow U.S. intelligence sources to apply statistical tools to the result of IE templating. In this manner, a consistent improvement in capability, interpreted by the advanced nature of information extracted in the templates, would be coupled with greater increasing incidence of subject related data within sources targeted by the U.S. IE tools. IE templates would begin to unwittingly identify these hoax references and establish a database of that country's nuclear capability. Additional other physical deception techniques could also be undertaken in conjunction with these methods to corroborate toward the purpose of creating an undetected deception.

Threat analysis for IE is premature, because the base technology has not yet fielded a highly capable system. However, it is relatively easy to imagine the weaknesses the United States presents to a potentially sophisticated adversary seeking to employ information warfare techniques and counter-extraction

[70] Another approach would be to "spoof" the collection apparatus into believing that the data source actually originated from Reuters, Wall Street Journal etc. This could be done by modifying only parts of the content of an electronically transmitted document, keeping the original "header information" (that identifying originating data) and retransmitting.

techniques in the future.

CHAPTER 7

EXTENDING CONCEPTS FOR THE EMPLOYMENT OF INFORMATION EXTRACTION

The roles, missions and organization of the military are described in its doctrinal materials using a number of models[71] extending down to the lowest functional and organizational military component. These models can be analyzed to suggest requirements for the information that is needed to accomplish the goals of the military system the model represents. If organizational and process models are well defined, then extending them to characterize an information architecture, or information model, is not difficult. The military employs several information systems techniques, such as those used for work breakdown structures, to develop information architectures in a similar manner. The result is a macroscopic projection of how military processes, organizations and functions interrelate, and what the information requirements are to support it.

Information extraction techniques can be developed to respond to information architectures that are created in the manner described above. By using a model-based approach for IE data can be extracted into templates created to meet continuous information requirements. Model based extraction can improve

[71] Models are abstractions of theories and systems. They are used to represent the properties of conceptual structures or the processes in systems that function toward a goal or product. Models of organizations and functional systems are beneficial in clarifying procedural responsibilities and interrelationships inherent within an organization's components. They can help to explain how and why complex objects are constructed or how sophisticated processes interrelate. Being able to "see" a visual representation of a problem often leads to its rapid solution.

the efficiency of IE systems by achieving a stable template design which is well integrated with the requirements of an information model. Representative models may be based on the National Strategic Decision Making model, those used for operational maneuver, or the performance steps for a specific occupational task.[72]

In summarizing these thoughts for model-based extraction, the reader should understand that a well developed process model must be available. This model must then be extended to identify information requirements that correspond to the processes. This is a different scope than that pursued in Brooks' concept for model-based retrieval which I discuss below.

Model-Based Retrieval and Model-Based Extraction

William Brooks, of Systems Research and Applications Corporation (SRA), presented a discussion of "Model-based retrieval: The Next Step Beyond Information Retrieval", at the Advanced Information Processing and Analysis Symposium, held in March 1995. His research hypothesizes a solution to the problems intelligence analysts have in fusing data that may be obtained through automated retrieval systems.[73]

Recall from a previous section that extraction technology creates information from data. Information retrieval technology is used to identify relevant sources that may be relevant to a user's research or analytical needs. The user of a retrieval system must read the contents of the documents that are identified by the retrieval system to extract information. Extraction technology advances the scope of

[72] Such as those presented in flowchart form, used to guide an individual in the completion of his assigned responsibilities.

[73] From an abstract concerning Model-based retrieval (MBR) research, Brooks, Scott Bennett, and Aaron Temin describe their approach as providing "an analytical framework for data fusion. MBR helps analysts to express the entire problem they are trying to solve as patterns using model-based search criteria. These criteria allow analysts to express complex temporal spatial, and functional relationships between retrieved elements of data...With this support, analysts can quickly model, monitor, and assess developing situations, identifying information gaps (for collection) and projecting a situation's future course (for prediction).

automation capabilities over that provided by retrieval systems by automatically yielding information from a source.

Model-based retrieval is employed to identify a set of available documents that pertain to a specific subject area. The goal of model-based retrieval is to help build the conceptual model (identify the problem) that represents the nature of the information query. The model may be identified through statistical methods, such as the comparison of the number of documents related to the subject. Defining the model becomes one of the products of the retrieval process. Therefore, the goal of an model-based retrieval system is identifying the problem. In this capacity, model-based retrieval has the potential to be beneficial in information and intelligence problems based in conceptual uncertainty and ambiguity.

In a complimentary way, model-based extraction can be applied to existing models that have well understood or continuous information requirements (in particular process models). The goal of model-based extraction is to make higher order, end-applications (command and control for example), function more efficiently.

Model-based extraction begins with a model, and in this sense assumes the characteristics of a knowledge engineering task.[74] The knowledge model is "solved" as if it were a series of problems, using IE capability. There are many opportunities to employ IE within the military through the concept of model-based extraction.

Hyperspace - A Frontier in the Quest for Relevant Information

The creation of a construct called *hyperspace* extends the concept for model-

[74] Knowledge engineering is the field of artificial intelligence that tries to use machines to help understand information. The problems attempted by a knowledge engineering system are usually presented as a conceptual or process model. The goal is for a machine to process information, traversing the model, until it makes conclusions concerning the relationship of the information with respect to the nature of the model. (Author's definition).

based extraction. A hyperspace is a domain of similar knowledge-based structures. It is established by interconnecting related information requirements that are shared amongst various government agencies within a relational information medium.[75] The concept of a hyperspace domain is a synthesis of ideas which include: model-based extraction; the planned interconnectivity features in the Defense Department's C4I Warrior program; the original goals for the Internet, the objectives of the new Intelink;[76] Davidow and Malone's ideas for the virtual corporation;[77] the opportunities that availed from the ability to link data using commercial-off-the-shelf hypermedia software, and the templating methods of IE.

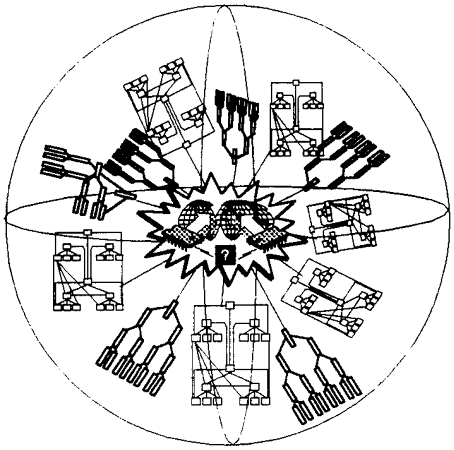


Figure 4 - A Hyperspace

[75] My search of the literature did not retrieve any reference to the term "hyperspace", and therefore I coin its definition here.

Examples of knowledge-based structures: process models, functional organizations, decision support systems.

[76] Douglas Waller, "Spies in Cyberspace", Time, March 20, 1995, pg 63-64. Intelink: "a worldwide computer network that has borrowed much of its technology from the Internet, the global network that links universities, research institutions...Intelink has 35 intelligence organizations feeding it so far...", operating at multiple levels of security classification.

[77] William H. Davidow and Michael S Malone, "The Virtual Corporation", New York: Harper Collins, 1992, pg 64: "The creation of the virtual corporation will result from linking relevant data bases into ever more extensive and integrated networks."

The prerequisite elements in building a hyperspace are a communications network, hypermedia software, and automated IE capability.[78] The only physical connection to the hyperspace occurs through the telecommunications hardware. A hyperspace is composed of a distinct set of knowledge-based structures that could represent information architectures, decision support systems, or organizational hierarchies. The hyperspace is represented by a sphere containing the information medium and the information requirements which are represented as hierarchical structures (author's creative limitation). The locus of the spherical representation I have drawn may be either a functional similarity, such as defense and national security interests, or a major grouping of common information requirements.[79]

Nodes and Leaves. The nodes inside the sphere are drawn to represent organizational hierarchies or conceptual models. The leaves of each node are then steps in a process, functional levels within an organization, or information template structures. Leaves are networks information requirements that could also represent a series of conditional statements (questions) that satisfy a particular requirement of the node.

Another aspect of the hyperspace concept is that nodes could be considered models of problems. The information needed to solve the node's problems is obtained automatically, either through the nodes organic IE capability, or acquired through links to other nodes and other extraction systems. Similar to the process used in a decision tree, the traversal of leaves and nodes then represents a

[78] Hypermedia/hypertext were terms coined by Theodore H. Nelson while attending graduate school at Harvard in the early 1960s, to describe a nonsequential archiving and retrieval method.

[79] A common information requirement exists between the Commerce Department and DoD in monitoring aspects of the defense industrial base, for example.

solution: Once all of the information requirements that make up a leaf are satisfied, the leaf's state is then determined once all of the leaves are known, the node is "solved."

How a Hyperspace Might Work. First, information extraction tools are employed to isolate information for specific needs. When data is extracted to meet a requirement it is "tagged" by the hypermedia system creating a conceptually linked network within the information environment. The properties of a hypermedia environment, developed in the systems design, would automatically establish contextual links to other pre-existing relevant data.[80] The logical links make information easier to retrieve by the other nodes in the hyperspace. It also ensures that relevant data is captured for the hyperspace, since it is possible that some template designs, used by some nodes, may miss some data. The logical design of a hyperspace makes it possible to conceive that the information requirement of one node might be a knowledge product of one of the other resident nodes.

The extraction of new data from the environment re-enforces the hyperspace concepts: relationships between existing data are reviewed against the newly extracted data; new relationships may be established between new data and old data; this cyclical process may identify relationships between concepts that were not previously known. Continuous analysis of hyperspace link activities between data and nodes may create new concepts for operational improvements and the further development of the hyperspace.

Roots in Several Spaces. Some knowledge acquisition requirements, and many

[80] The links may be directly designed and established by the analyst of a dataset, or user of the information, employing hypermedia authoring system, or indirectly defined or inferred by use of pattern or string (as in a string of letters or words) matching and artificial intelligence (AI) techniques, such as those found in the information extraction niche disciplines.

organizations require fluidity in purpose and flexibility of response in order to maintain their effectiveness. For example, the military's special operations units have structures and requirements spanning a wide range of military and civil functions. The complexities embodied by these information requirements may not be reduced to one fixed set of information models, and therefore these organizations may maintain nodes in several hyperspaces.

Advantages. The hyperspace shares remote interconnected resources in satisfying information requirements that may originate from any member, or "node," of the hyperspace. Efficiencies are gained from the pooling of information resources, the processing of similar or identical information requirements collectively, and minimizing aggregate software maintenance requirements.[81] By creating a logical relationship between information requirements, rather than between organizations, objectivity in information processing can be achieved.

A hyperspace can serve to eliminate cultural impediments to processing and distribution of information by focusing at the requirements level rather than the organizational level. Certainly, organizations are the center of information requirements through their roles and function within government. However, many redundant architectures are employed in obtaining information that may serve similar purposes across organizational boundaries (and the idea for voluntary sharing of information in the spirit of "teamwork" doesn't work). In a hyperspace, dissemination, or distribution of information, is replaced by the inter-linking of organizations through their requirements, and their databases.

[81] Software maintenance is generally estimated to consume up to seventy percent of a systems life-cycle costs.

CHAPTER 8

CONCLUSION

. IE technology is now approaching viability as a functional military capability, and its potential is limited only by the lack of military applications to direct its development. Hence, specific defense purposes for automated information extraction technology should be developed. To assist in attaining this objective, deliberate efforts should be undertaken to resolve the meaning of IE terms, and to clarify abstract concepts. A discipline for this field of research should be formalized to promote its general understanding across the range of functional defense concerns. This research standardization could serve to stimulate greater interest and cooperation within the military for exploring opportunities to exploit IE technology.

Technical Complexity

Clearly extensive academic background is necessary to gain a deep understanding of complex issues concerning advanced information processing technology. However, uses for information extraction technology can be developed with only a rudimentary background in information and computer science, combined with a review of military intelligence doctrine.

Developmental Outlook

Natural language and image processing technology are disciplines rich in potential for military application. However, many technical issues must be overcome in a wide range of related technical disciplines. Realistically, effective information extraction products will not be fielded in the near future.

Practical application of even the most fundamental, text-only, IE systems is perhaps fifteen years away. There are many systems in development, and it is possible that some early demonstration systems will be fielded within the next five years. However, accurate and reliable IE capability must overcome significant performance problems in areas that may not be adequately resourced. Along with these IE-specific component and systemic technical problems, IE will not be practical until knowledge-based applications are developed that can interface with the output from IE systems.

Academic Thrust

Information extraction technology applies knowledge from such fields of information and computer science, engineering, mathematics, linguistics, psychology, cognitive science, philosophy and ergonomics. This listing of specialty areas which are making relevant and specific contributions to advanced information processing and artificial intelligence efforts suggests that information theory may belong inside of many academic disciplines and technical specialties.

An evolution in the understanding of information sciences, human dynamics, and systems management, has expanded the scope of interest related to the potential capabilities for microprocessor-enabled automation of intelligence[82] tasks. The serious interests in advanced information processing technology from

[82] Academic sense versus "military intelligence."

seemingly distant academic and professional career fields may suggest the creation of new career fields which will require expertise in both hard (i.e.-software and electrical engineering) and soft (cognitive and social sciences) skills.

New Approach to Cognitive Sciences?

The use of context vector technique for interpreting both text and imagery data has been mentioned in the literature, highlighting the progress being made by HNC Software, Inc.[83] The use of a common set of mathematical techniques in the processes for text extraction and image understanding, demonstrates a new perspective on the problem of transferring human sensory interpretation abilities to machines. If the same core computational algorithms can be proven to be valid in representing two distinctly different abstractions of visual sense, then perhaps a better understanding of the cognitive human skills required for other intelligent processes can be gained from this research area.

[83] Discussed in a conversation with William R. Caid, Director, Information Retrieval Systems, HNC Software, Inc., during the AIPA Symposium on 29 March 1995.

GLOSSARY

Anti-Information - Potentially relevant information that is not processed due to resource or policy issues which ultimately increases the level of uncertainty in the decision cycle. Anti-information can be symptomatic of strategic design problems in information support systems or the hesitancy in organizations to integrate information quickly.

ASCII - American Standard Code for Information Interchange; the common binary code for the English alphabet character set.

Concept - A psycho linguistic category that can be either physical (weapons or divisions) or abstract (attitude or politics).

Data Mining - The unspecified techniques applied to the finding of human concepts amidst huge amounts of data

DBMS - Database Management System.

Digitization - Conversion of one data form into binary code. Specific to speech recognition the digitization process converts analog voice into an electrical signal that can be represented by 1's and 0's.

Document - Anything recorded (written, printed, taped, image) which may serve to prove something, or record an event, or describe an activity.

Document Detection - The capability to locate documents containing the type of information the user wants from either a text stream or a store of documents. text retrieval and message routing (dissemination) are both types of document detection. Often used synonymously with "information retrieval."

Hyperspace - A domain of similar knowledge-based structures established by interconnecting their related information requirements, or templates. The principle communication method is through the linking of data relationships between conceptual models.

Image Base - A database of images (paper documents, photos, video clips, graphics) that is indexed by some key header information, such as last name or social security number, date-time group, to facilitate retrieval upon request.

Imagery - Data collected that generates the visual representations of spatial objects. Satellite photographs, photos from still cameras, video recordings, X-rays, and radar systems, among other data forms, are considered imagery.

Information Extraction (IE) - The capability to locate specified information within a text (ARPA definition).

Information Retrieval (IR) - Identifying sources of information corresponding to a subject and related concepts. Similar to a search in a library.

Machine Translation (MT) - Automated language conversion through intelligent machines.

Model-based Retrieval (MBR) - The building of problem or concept models through statistical analysis of information retrieval performance results over a given subject domain.

Model-based eXtraction (MBX) - Solving problems that can be represented as information models using automated information extraction.

NLP - Natural Language Processing. The field of artificial intelligence research that seeks to teach machines how to interpret human language presented in text format.

Bibliography

Proceedings

Advanced Information Processing and Analysis Steering Group, Advanced Research and Development Committee, "Symposium on Advanced Information Processing and Analysis" 28-30 March 1995. (Symposium reference book) n.p., 1995.

Advanced Research Projects Agency Software and Intelligent Systems Technology Office. Proceedings - Fifth Message Understanding Conference (MUC-5). San Mateo, CA: Morgan Kaufmann, 1993.

Advanced Research Projects Agency Software and Intelligent Systems Technology Office. Proceedings of the Human Language Technology Workshop. San Francisco: Morgan Kaufmann, 1994.

Advanced Research Projects Agency Software and Intelligent Systems Technology Office. Proceedings - TIPSTER Text Program (Phase I). San Francisco: Morgan Kaufmann, 1993.

Defense Advanced Research Projects Agency Software and Intelligent Systems Technology Office. Proceedings - Third Message Understanding Conference (MUC-3) San Mateo: Morgan Kaufmann, 1991.

Defense Advanced Research Projects Agency Software and Intelligent Systems Technology Office. Proceedings - Fourth Message Understanding Conference (MUC-4). San Mateo, CA: Morgan Kaufmann, 1992.

Hirschman, Lynette. "An Adjunct Test for Discourse Processing in MUC-4". Proceedings - Fourth Message Understanding Conference (MUC-4). San Mateo, CA: Morgan Kaufmann, 1992.

Hobbs, Jerry. "The Generic Information Extraction System." Proceedings of the Fifth Message Understanding Conference (MUC-5). San Francisco: Morgan Kaufmann Publishers, Inc, 1993.

Hobbs, Jerry, and Israel, David. "Principles of Template Design", Proceedings of The Human Language Technology Workshop, March 8-11, 1994. San Francisco: Morgan Kaufmann Publishers Inc.

Okurowski, Mary Ellen. "Information Extraction Overview", Proceedings - TIPSTER Text Program (Phase I). San Francisco: Morgan Kaufmann, Sept. 1993,

Onyshkevych, Boyan. "Template Design For Information Extraction". Proceedings from the Fifth Message Understanding Conference (MUC-5) San Francisco: Morgan Kaufmann Publishers Inc, 1993.

__________________. "Issues and Methodology for Template Design For Information Extraction", Proceedings of The Human Language Technology Workshop, March 8-11, 1994. San Francisco: Morgan Kaufman.

Will, Craig A. "Comparing Human and Machine Performance for Natural Language Extraction: Results for English Microelectronics from the MUC-5 Evaluation", Proceedings of the Fifth Message Understanding Conference (MUC-5). San Francisco: Morgan Kaufmann Publishers Inc., 1993.

Books

Andriole, Stephen J., and Hopple, Gerald W. Defense Applications of Artificial Intelligence - Progress and Prospects. Lexington, MA: DC Heath and Company, 1988.

Davidow, William H., and Malone, Michael. The Virtual Corporation-Structuring and Revitalizing the Corporation for the 21st Century", New York: Harper Collins, 1992.

Handel, Michael I. Intelligence and Military Operations. Portland, OR: Frank Cass and Company Limited, 1990.

__________________. War, Strategy, and Intelligence. London: Frank Cass and Company Limited, 1989.

Lehner, Paul E. Artificial Intelligence and National Defense - Opportunity and Challenge. Blue Ridge Summit, PA: TAB Professional and Reference Books, 1989.

Minsky, Marvin (Ed.) Semantic Information Processing. Cambridge, MA: The MIT Press, 1968.

Patterson, Dan W. Introduction to Artificial Intelligence and Expert Systems. Englewood Cliffs, NJ: Prentice-Hall, 1990.

Richeson, Jeffrey T. The U.S. Intelligence Community (2nd Ed.). New York: Harper-Collins, 1989.

Salton, Gerald. Dynamic Information and Library Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Tucker, Allen B., Jr. Programming Languages (Second Ed.) New York: McGraw-Hill, 1986.

Periodicals

Kelly, Brian J. "Command and Control Basics Enhance Explosion of Information Technology." Signal, March 1995, pp. 15-16.

Lehnert, Wendy and Sundheim, Beth. "A Performance Evaluation of Text-Analysis Technologies." AI Magazine, Fall 1991, pp. 81-94.

Waller, Douglas. "Spies in Cyberspace." Time, March 20, 1995, pp. 63-64.

Wilson, Patrick. "Unused Relevant Information in Research and Development." Journal of the American Society for Information Science vol. 46, no. 1, 1995, pp. 45-51

General Sessions of the AIPA Symposium, 28-30 March 1995

James Foley, Director of Graphics, Visualization and Usability Center, Georgia Institute of Technology. "Information Visualization: The Next Frontier for Computer Graphics." 28 March 1995.

Leo Hazlewood, Executive Director, Central Intelligence Agency. A general address focusing on advnced information processing technology and concerns for insuring that end-user needs are being met. 28 March 1995.

Kenneth A. Minihan, MG, U.S. Air Force Assistant Chief of Staff, Intelligence. "Future Challenges Technology Templating", 29 March 1995.

Thomas R. Pedtke, Technical Director of Data Exploitation, National Air Intelligence Center. "The Open Source Information System & Textual Information Process Reengineering." 30 March 1995.

Edward Thompson, Director, ,Software and Intelligence Systems Technology Office, Advanced Research Projects Agency (ARPA), "Enabling the Intelligence Analyst to Do More With Less in a Volatile World", 29 March 1995.

Seminar Presentations at AIPA Symposium, 28-30 March 1995

HNC Software Inc. "Docuverse: Graphical Representation of Information Using Context Vectors", HNC, 5930 Cornerstone Court West, San Diego, CA

Mitre Corp. "Evaluating Techniques for Integrated Data Access", a presentation by Dr. Len Seligman and Dr. Marcia Kerchner.

National Drug Intelligence Center (NDIC), "The NDIC Organizational And Strategic Intelligence System-OASIS"

Papers Presented at the AIPA Symposium, 28-30 March 1995

Brooks, William and others. "Model-based Retrieval: The Next Step Beyond Information Retrieval." Systems Research and Applications Corporation, 2000 15th St. North, Arlington, VA. 22201.

Caid, William R., and Carleton, Joel L. "Context Vector-Based Text Retrieval." San Diego: HNC, Inc.

Hendrickson, Timothy B. "Exploring the Galaxies. Applying Visualization to Analysis of Textual Information."

Kim, Yeongji, Melvin, Bill and Bobick, Aaron. "Computer Vision Tools for the Exploitation of Video Imagery"

Sasseen, Robert V., and Caid, William R. "Docuverse: Graphical Representation of Information Using Context Vectors." San Diego: HNC, Inc.

Reports and Documents

Advanced Information Processing and Analysis Steering Group (AIPASG). "P1000 Strategic Plan for Information Visualization." n.p. 1994.

Advanced Research Projects Agency Software and Intelligent Systems Technology Office. "TIPSTER Text Phase I 24-Month Conference Executive Review." n.p., 1993.

Breault, Holly M. and Jumper, Eric J., Jr. "Evaluation of the Advanced Reasoning Theory Message Understanding System RL-TR-94-88, Griffiss AFB, NY: Rome Laboratory, 1994.

Logicon, Inc. "Message Understanding For Automated Data/Knowledge Base Update." RADC-TR-89-245, Vol I (of two), Griffiss AFB, NY: Rome Air Development Center, 1989.

Sundheim, Beth "Second Message Understanding Conference (MUCK-II) Report." Naval Ocean Systems Center Technical Report 1328, September 1989.

Unpublished Papers

Fox, Steven G. "Technology's Limiting Effect Upon Intelligence Politicization." U.S. Naval War College, Newport, RI: 1995.

_____________. "Unintended Consequences of Joint Digitization." U.S. Naval War College: 1995.

James, Glenn E. "Chaos Theory: The Essentials for Military Applications" Unpublished Research Paper, U.S. Naval War College, Newport, RI: 1995.

Military Doctrinal Publications

Joint Chiefs of Staff. "National Military Strategy of the United States of America, A Strategy of Flexible and Selective Engagement." February 1995.

U.S. Dept of Army. Information Operations (Coordinating Draft), FM 100-6. n.p.: 22 July 1994.

U.S. Dept of Army, Training and Doctrine Command. Force XXI Operations. TP 525-5.

U.S. Dept of Navy. Naval Intelligence, Naval Doctrinal Publication 2. n.p. September 1994.

U.S. Dept of Defense. Intelligence Support for Operations Joint Publication 2-0. U.S. Government Printing Office, 1993.

Interviews and Telephone Conversations

Interview with William R. Caid, Director, Information Retrieval Systems, HNC Software, Inc., while attending the AIPA Symposium in Tysons Corner, VA, on 29 March 1995.

Interview with Ms. Anita Cohen, Chair of AIPASG, during AIPA Symposium on 28 March 1995.

Interview with Dr. Louise Guthrie of Lockheed Martin during AIPA Symposium on 28 March 1995.

Interview with MAJ Joseph L. Hollett, Rand Research Fellow, Project Air Force, Santa Monica, CA while attending the Information Warfare Conference held at the Naval War College, 3 March 1995.

Interview with Dr. Marcia D. Kerchner and Dr. Leonard Seligman of Mitre Corporation, while attending the AIPA Symposium in Tysons Corner, VA, on 29 March 1995.

Interview with LTC William Paul, Director of Information Management Branch, Combat Applications Group, Fort Bragg NC on 18 April 1995.

Interview with LTC Steve C. Schrum, Intelligence Officer, Combat Applications Group, Fort Bragg, NC on 14 April 1995.

Interview with Ms. Sarah Taylor, Chair of ARPA TIPSTER architecture committee during AIPA Symposium on 28 March 1995.

Telephone conversation with MAJ Robert M. Evans, senior analyst, Army Information Warfare Office, 16 March 1995

Telephone conversation with Rebecca Davis, Senior Analyst, Pathfinder Project, National Ground Intelligence Center, Charlottesville, VA. 14 March 1995.

Telephone conversation with Tim B. Hendrickson, Manager, Pathfinder Project, National Ground Intelligence Center, Charlottesville, VA. 15 March 1995.

Telephone conversation with Sharon K. Kaufmann, Digital Systems Research, Inc., Arlington, VA. 16 March 1995.

Other Sources

Sterling Software. "The CDIS Automatic Templating System (ATS) vs. Manual Templating", a memorandum prepared by Richard Lee and others. Sterling Software, Language Analysis Systems, Chantily VA, 25 August, 1994.

Conquest Software Inc. "NetQuest: Natural Language Access to Distributed, Heterogenous Information Sources"

Central Imagery Office "Providing Customers Worldwide Imagery Services on Demand...the United States Imagery System." 8401 Courthouse Road, Vienna, VA 22182-3820

Sietec Language Technology. "METAL: The Global Solution." 113 A Leslie St., Don Mills, Ontario, Canada. Product literature that describes generic machine translation processes.

Turabian, Kate L. A Manual for Writers (Fifth Ed.) Chicago: University of Chicago Press, 1987

U.S. Naval War College. "Style Manual and Classification Guide." Newport, RI: 1995.

Electronic Documents

Lehnert, Wendy. "Information Extraction", Internet WorldWide Web address at http://www.cs.umass.edu

Los Alamos National Laboratory, "Data Mining for Concept Extraction." through home page at http://www.lanl.gov.